"Hoovered up as a data point": Exploring Privacy Behaviours, Awareness, and Concerns Among UK Users of LLM-based Conversational Agents

Lisa Mekioussa Malki
University College London
lisa.malki.21@ucl.ac.uk

Akhil Polamarasetty
University College London
akhil.polamarasetty.23@ucl.ac.uk

Majid Hatamian
Google
hatamian@google.com

Mark Warner
University College London
mark.warner@ucl.ac.uk

Enrico Costanza
University College London
e.costanza@ucl.ac.uk

Abstract

Large Language Models (LLMs) are widely used in conversational agents (CAs) due to their ability to generate coherent and human-like text. However, their deployment raises significant privacy concerns, as users often share sensitive data in prompts. This data can be used to train the underlying LLM, introducing memorisation risks and challenging users' right to be forgotten. While these issues have been explored from a technical standpoint, little is known about how users perceive and navigate privacy issues in their day-to-day use of LLM-based CAs. To address this research gap, we conducted a survey of UK-based CA users (n = 211) that focused on their privacy behaviours, self-disclosure boundaries, concerns, and awareness of LLM-specific privacy issues. We found that engagement with protective behaviours was low overall, and that many participants held inaccurate beliefs about the effects of deleting data and opting out of model training. Although participants were generally reluctant to share sensitive information during interactions, we identified several challenges to limiting self-disclosure in practice, such as balancing privacy with app utility. Lastly, we observed a nuanced relationship between privacy awareness and concern, and identified significant demographic effects. We propose design avenues for privacy-supportive tools, and discuss the implications of our work for regulation and governance.

Keywords

Large Language Models (LLMs), AI, conversational interfaces, chatbots, usable privacy, human-computer interaction

1 Introduction

LLMs are trained on vast amounts of textual data, allowing them to perform well on several natural language processing tasks [24]. Motivated by their strong performance and versatility, LLMs have been widely deployed in CAs such as OpenAI's ChatGPT, which users can interact with via text or voice [118]. Compared to CAs with specific use-cases (e.g., customer service chatbots), it is difficult to anticipate the types of data that apps like ChatGPT will collect, as they support multimodal inputs and can converse on a wide range of topics [91]. Growing evidence suggests that users include personal data in their prompts to support their tasks—for instance, sharing their name and employment history when drafting a CV [77, 114] or uploading confidential workplace material for analysis [60]. Owing to their sophistication, LLM-based CAs are capable of employing persuasive strategies and exhibiting anthropomorphic characteristics, increasing the likelihood that users will disclose more personal information than they intend [47, 57, 102].

The risks of disclosing sensitive data to LLM-based CAs go beyond traditional privacy concerns such as unauthorised collection or sale of data [114]. Chat data is often used to train and fine-tune the underlying model, which can lead to the unintentional memorisation of this data [60]. This not only raises the risk of data extraction [20], but also poses compliance challenges regarding users' right to be forgotten, since it is difficult to selectively remove memorised information from a trained model [90, 111]. Growing evidence suggests that users lack awareness of these risks and misunderstand how LLMs operate on a technical level, limiting their ability to give informed consent and protect themselves against privacy risks [35, 61, 114]. Further, while technical mitigations such as training data sanitisation, differentially private training, and machine unlearning are fast-developing, these techniques have yet to be scalably applied to LLMs in industry, and currently fail to address the rich and context-dependent nature of user self-disclosure to CAs [16, 69]. Against this background, it is important to examine how users understand and manage privacy risks related to LLM-based CAs, with a view to developing user-centered privacy interventions.

As it stands, most existing LLM privacy research is technical and model-centric, focusing on evaluating data leakage risks [7, 20, 44, 81] or developing technical mitigations [90]. While valuable, model-centric research provides little insight into end-users' understanding of privacy risks and protective strategies [114]. Similarly, while conversational interfaces have been extensively studied in the field of usable privacy (e.g., [22, 23, 57]), most of this work considers rule-based chatbots or explores user privacy psychology in a way that is agnostic to the chatbot's underlying language model. The use of LLMs in CAs introduces a distinctive threat landscape characterised by opaque, complex training processes and multimodal capabilities. However, dedicated user research on LLM-based CAs remains scarce, and is limited to smaller-scale qualitative interview

Proceedings on Privacy Enhancing Technologies 2025(4), 838–860

https://doi.org/10.56553/popets-2025-0160

studies [61, 114]. To address the lack of wider LLM-focused usable privacy research, we conducted an online survey study examining privacy behaviours, awareness, and concerns among regular users of popular LLM-based CAs. We addressed the following research questions:

RQ1 What privacy behaviours do users engage in when using LLM-based CAs?
RQ2 What types of information are users willing to share with LLM-based CAs, assuming this information is relevant to their tasks?
RQ3 What are users' expectations of data deletion, retention, and opting out of having their data used to train models?
RQ4 How aware are users of the privacy issues associated with LLMs such as model training, data memorisation, and data extraction, and how concerned are users about these issues?
RQ5 What challenges do users face when trying to protect their privacy when using LLM-based CAs, and what solutions or strategies could support them?

Overall, we found engagement with common privacy-protective behaviours to be low (***RQ1***), with just 7% opting out of their data being used to train models. While participants had clear boundaries for self-disclosure that broadly matched the sensitivity of the data being shared, many struggled to balance privacy with prompt utility, and were concerned that inferences could be drawn about them without their knowledge (***RQ2***). Further, many participants held mismatched expectations of opting out of model training and data deletion, and believed that information could be retroactively unlearned from the model (***RQ3***). While participants perceived several privacy practices and risks to be hypothetically plausible, they had lower baseline awareness and expressed varying levels of concern about them (***RQ4***). Contextualising these findings, our qualitative analysis uncovered feelings of privacy cynicism, overwhelm with the technical complexity of AI, and practical barriers to privacy action. Addressing these barriers requires better privacy interaction design, educational strategies, and regulation (***RQ5***). Finally, our findings suggest that demographics including age, gender, education, and reasons for using LLM-based CAs, likely influence users' privacy attitudes and behaviours. Our study makes the following novel contributions:

- We present a novel mixed-methods usable privacy study of LLM-based CAs covering privacy attitudes, knowledge, and behaviours. Our study illustrates a complex interplay between privacy strategies, awareness, and concerns that is shaped by demographics and contexts of app usage.
- We present findings from a UK context, providing a baseline for comparison with future work that explores cross-cultural privacy perceptions of generative AI.
- We identify several privacy challenges faced by users including privacy cynicism, difficulties in calibrating information disclosure, and a limited understanding of LLM training and inference.
- We identify diverse avenues for future research and design, such as novel privacy choice architectures for CAs and semi-automated support for prompt sanitisation.

2 Literature Review

We now summarise related literature. This section examines the technical privacy risks associated with LLMs (§2.1), and how these risks translate into privacy compliance challenges when LLMs are integrated into CAs (§2.2). We then explore key concepts in usable privacy, including the privacy paradox, privacy calculus, and the impact of demographics on privacy attitudes and behaviours (§2.3).

2.1 Privacy Risks of LLMs: Inference Attacks, Data Extraction, and Memorisation

Large Language Models (LLMs) use an attention-based transformer to process, understand, and generate natural language. Their impressive performance across various tasks has driven their adoption in conversational interfaces, outpacing traditional rule-based chatbots and smaller language models [105]. However, LLMs have well-known privacy shortcomings: the extensive corpora required to train them can include personal data [108], and privacy vulnerabilities pervade the training and inference pipeline. Membership inference attacks allow adversaries to determine whether data pertaining to a specific individual was used to train an LLM [88], and the closely related property inference attack is concerned with extracting global properties of the training dataset [115]. Recently, Wang et al. demonstrated a novel property existence attack, using shadow models and similarity computation to ascertain whether samples with specific properties (e.g., gender or ethnicity) were present in the training set of a generative model [106].

Inference attacks can precede data extraction attacks by acting as an oracle to identify target samples which have a high probability of being in the training set [16]. Data extraction from LLMs is made possible by *memorisation*, which refers to the ability of deep neural networks to recall and reproduce strings they have seen during training [60]. While memorisation occurs in all pre-trained language models, LLMs memorise a greater portion of training data, and do so more readily than smaller models [96]. For example, Carlini et al. [20] used membership inference to extract users' names, addresses, and social media profiles from GPT-2, and Bai et al. [7] demonstrated an increased risk of memorisation and data leakage for text containing special characters, such as *@* in email addresses.

2.2 Privacy Compliance in LLM-Based Apps

On an organisational level, sensitive information disclosed to LLM-based CAs can be retained for long periods of time, accessed by developers, and used to train and fine-tune models [46, 60, 83]. In addition to established security risks like cyberattacks and API misuse, LLM-based apps pose the novel technical risks summarised in §2.1, and publicly available, black-box models can be jailbroken to reveal user data that it has been trained on [50]. Moreover, the downstream uses of LLMs are difficult to predict, and memorisation risks are not well understood by lay users [16, 114], complicating the ability for users to meaningfully consent to having their data used to train models [16, 114]. This poses a compliance challenge for frameworks such as the GDPR, which depend on consent as a lawful basis for processing [111].

Furthermore, the permanence of data traces within LLMs makes the right to be forgotten difficult to fulfill [56, 60, 111]. Under Article 17 of the GDPR, a data subject has the right to have personal data

erased if it is no longer required, or if a legal basis for processing is altered (e.g., the user has withdrawn their consent) [56]. In practice, erasure should be complete and cover both live copies of data as well as backups. However, even if user data is deleted from the company's database and excluded from future training iterations, data traces can persist within model memory until it is retrained from scratch, which is computationally infeasible for each deletion request [111]. To address this challenge, machine unlearning has emerged as a way to remove specific, undesired data items from a trained model [69, 90]. However, unlearning techniques perform questionably when applied to LLMs with vast training datasets [69], and enforce rigid assumptions about the nature of private data which may not hold for natural language, such as it being well-structured and easy to remove [16]. For these reasons, no widespread industrial implementations of machine unlearning for LLMs currently exist [13, 16, 69, 112].

2.3 Usable Privacy and Conversational AI

Usable privacy research explores how users perceive the flow of their data through socio-technical systems using *privacy constructs*: specific aspects of users' privacy psychology which include concerns, trust, and self-disclosure [31]. Our study focuses on privacy awareness, concern, and behaviour. Privacy awareness is the extent to which users understand how their data is collected and used [45], while privacy concern reflects the worry or apprehension about situations that involve the transmission or use of their data [31]. Privacy behaviour refers to any observable action a user takes to keep their data safe, such as limiting disclosure, deleting data, and seeking information about privacy [12].

2.3.1 Privacy Awareness and Concern. Several studies have explored CA users' awareness and concerns about privacy risks and practices, revealing that users are primarily concerned about the collection and storage of chats by developers [9, 23, 61, 114], adverse events such as unauthorised data sharing and cyberattacks [2, 4], and the creation of personalised inferences or profiles [70, 116]. By contrast, LLM-specific risks such as model training and memorisation are not well-understood by users, and perceived as too abstract to warrant concern [114]. Zhang et al. [114] found that many participants held erroneous mental models of ChatGPT response generation (e.g., assuming that LLMs work like search engines), and that such misunderstandings hindered their ability to anticipate memorisation. Therefore, a shallow or incorrect understanding of AI systems can obscure potential risks, hinder users' ability to take protective actions, and worsen privacy concerns, since users lack concrete assurances that their data is being processed in ways that align with their expectations [116].

2.3.2 Privacy Behaviours: Self-Disclosure and the Privacy Paradox. Users of CAs employ various methods to protect their privacy, with the most common approach being to limit the amount of information they disclose during conversations. Antecedents of disclosure to CAs include anthropomorphic design cues [27, 57, 64, 98], trust in the system [28, 54], and concerns about data leakage [10, 23, 113]. Prior research has explored users' intended disclosures in hypothetical scenarios [10, 23, 70], as well as their actual disclosure behaviours through experimental studies and analyses of chat logs [77, 110, 114], often revealing a discrepancy between the two.

The nuanced interplay between privacy concern, awareness, and self-disclosure is well-understood within the usable privacy discipline. Despite being aware of privacy risks and expressing concern, users frequently neglect to adopt privacy-protective behaviors—a phenomenon named the 'privacy paradox' [62]. Observed inconsistencies between self-reported privacy attitudes and real-world behaviour have long been attributed to a 'privacy calculus' wherein the benefits of information disclosure (e.g., personalisation and convenience) are logically judged to outweigh the privacy risks [37]. There is evidence of this dynamic at play in CAs. For instance, Chalhoub et al. [22] found that Amazon Alexa users preferred their devices to be 'always-on' to reduce the manual burden of activating them, and ChatGPT users often include sensitive and personal data in prompts for convenience and increased app utility [114, 116].

At the same time, it is well-known that user disclosures are influenced by external factors. For example, while design nudges can be used to encourage privacy-conscious disclosures [32, 65], the same dynamics can be abused: advanced CAs can manipulate users into prolonged app usage, or encourage deeper self-disclosures through invasive questioning [3, 49, 87]. User distrust in service providers, combined with a lack of meaningful transparency and limited privacy controls, can lead to privacy cynicism and disengagement [53]. This often results in resigned inaction that may be mistaken for apathy [39]. Therefore, rather than attributing the privacy paradox solely to users prioritising convenience over privacy, it should be understood in the context of broader socio-technical power imbalances [39, 114].

2.3.3 Demographic Influences. Finally, several studies have explored the effect of demographics—particularly gender, age, and education level—on self-disclosure, privacy awareness, and concern. For instance, Belen-Saglam et al [10] ranked several information types by perceived sensitivity and willingness to disclose among a UK sample. Gender emerged as a significant mediator of privacy perceptions, and older adults rated a higher proportion of information types as sensitive compared to younger age groups. Other studies have similarly reported heightened concern among older adults regarding data deletion and misuse by CAs [9, 35].

However, in a healthcare context, older adults have expressed lower levels of privacy concern, owing to a higher level of trust in chatbots associated with official healthcare providers [41, 70]. Education level has also shown mixed effects, with some studies finding it increases privacy concern and awareness and others observing the opposite depending on the specific technology being evaluated [9, 12]. It is clear that demographic factors do not uniformly influence privacy attitudes; their effects vary depending on the specific technology and context, highlighting the relevance and timeliness of our analysis of UK users of LLM-based CAs.

2.4 Study's Novelties

Our work builds on related studies in several important ways. Firstly, the scope of our work is novel compared to prior experimental usable privacy studies of CAs which focus on individual privacy psychology, and model single constructs like concern or self-disclosure in isolation [28, 29, 41, 57, 66]. These studies also

focus primarily on rule-based chatbots, and even those that address ChatGPT (e.g., [29]) do not engage with the distinct technical privacy risks posed by LLMs such as model training and memorisation. By contrast, our exploratory study covers psychological, organisational, and technical dimensions of LLM privacy, bridging the gap between technical and user-centric privacy literature.

Against this background, the studies most similar to our own are mixed-methods user surveys that examine privacy behaviours, concerns, and awareness regarding AI chatbots for mental health [23] and LLM-powered healthcare consultations [70]. Our methods and sample differ from these studies. In addition to focusing on healthcare, the participants in these studies were overwhelmingly not current LLM users, making the findings more relevant to the general public than to users of LLM-based CAs. Our study also explores interface design, and incorporates mockups for probing participants' mental models of specific privacy features, and developing learnings for privacy design. Lastly, our work builds on qualitative user studies of ChatGPT [61, 114] by gathering data on a larger scale, in a different cultural context, and analysing the effects of demographics and contexts of app usage.

3 Methods

3.1 Recruitment

To address our research questions, we administered an online survey to 211 participants via the academic recruiting platform Prolific. We used a screening survey to recruit participants that were aged over 18, lived in the UK, and used one or more LLM-based CAs at least monthly. We recruited UK residents to control for international differences in app access and cultural privacy attitudes [10, 68]. We selected monthly use as a lower bound for participation to ensure a baseline level of familiarity with the technology while still capturing a range of usage patterns. Unlike previous user studies that focus solely on ChatGPT [61, 114], we recruited users of any LLM-based CA to improve the generalisability of our study, and capture a wider range of user interactions with AI. We did not mention privacy in our study title, description, or screening survey to avoid priming participants or skewing our sample [14, 114]. Participants were paid GBP0.13 to complete the screening survey and GBP1.25 for the full survey which took between 5 and 10 minutes to complete. Of the 375 participants who completed the screener, 221 completed the full survey. Ten responses were discarded due to failed quality checks, leaving 211 responses for analysis.

3.2 Survey Instrument

We designed our questionnaire in three phases: identifying target constructs, developing an initial item pool, and iterative piloting [52]. Due to the lack of user studies on LLM-based CAs, we explored a broad range of constructs including privacy behaviours, self-disclosure intentions, comprehension of privacy features, and privacy awareness and concern. To ground our findings in users' experiences, we asked open-ended questions about challenges participants faced when maintaining their privacy while using LLM-based CAs and potential solutions. For the full survey questionnaire, see Appendix A.

3.2.1 Privacy Behaviours. Our first research question was to explore users' self-reported privacy behaviours when interacting with LLM-based CAs. To identify behaviours of focus, we reviewed diverse usable privacy research on CAs [23, 114], general internet use [18, 31, 86], and fitness apps [45]. We identified six common privacy behaviours which are summarised in Table 1. These behaviours were slightly modified to be relevant to the most popular LLM-based CAs used by participants (ChatGPT, Google Gemini, and Microsoft Copilot). For example, we added using of data export features to our list of behaviours, and adjusted the generic 'opting out of secondary use' behaviour to opting out of model training specifically. Privacy behaviours were queried in a retrospective assessment format, with participants indicating which they had previously engaged in from a provided list. To add further context, we also asked participants whether they were aware of the following privacy features before completing the survey: chat deletion, export, and opting out of having chats used for model training.

Table 1: Privacy behaviours adapted from previous literature.

Behaviour	Citations
Reading privacy policies	[18, 45]
Limiting disclosure of sensitive information	[23, 31, 86, 114]
Data falsification (signing up with a fake email or pseudonym)	[23, 86, 114]
Opting out of secondary uses of data	[18, 31, 45, 86]
Using privacy tools (e.g., VPNs)	[18, 23, 31, 86]
Deleting data	[23, 86]

3.2.2 Information Disclosure Boundaries. We measured participants' information disclosure boundaries by assessing their willingness to share 12 data types during a hypothetical interaction with an LLM-based CA. Participants rated their willingness on a five-point Likert scale, where five was the most willing. Here, we considered two core antecedents of online disclosure to CAs: task relevance and information sensitivity. We addressed task relevance by instructing participants to assume that each type of information was relevant to their task. We developed 12 data items of varying sensitivity levels using prior related work as a baseline [10, 21, 23, 72, 78]. We extracted the information types explored in each of the above studies and removed duplicates, resulting in 42 granular data items, which we list in Appendix B.

To avoid participant fatigue, we reduced the 42 items into 12 information categories based on topical relationships and sensitivity. For example, prior work suggests that UK users are particularly concerned about health and financial data [10], motivating us to separate these attributes into two categories, allowing a granular analysis of users' perceptions. Similarly, we created two demographic categories: basic information (e.g., age, gender) and information considered sensitive under the GDPR (e.g., religion and sexual orientation). We created a separate category for PII a user might disclose during account creation, such as full names and emails. Finally, we included items not covered in previous studies such as work documents and photos, to reflect the multi-modal capabilities of LLMs.

3.2.3 Comprehension of Privacy Features. Addressing our third research question, we explored participants' comprehension of three tasks: opting out of model training, chat history deletion, and account deletion. We focused on these features as they are implemented in at least one CA used by participants and have distinct, and frequently misunderstood, outcomes [92, 97, 114]. To enhance realism, we presented participants with images of each feature. Rather than using screenshots of existing apps, we developed fictional mockups for the survey. These mockups were designed to be familiar to participants regardless of which CAs they used, and excluded any branding or company names to avoid influencing their judgments. Our mockup images are provided in Appendix A.

We developed mockup screens of a main chat window, a summary of the user's chat history, and a privacy settings screen containing three options: opting out of model training, chat history deletion, and account deletion. The design of our mockups was inspired by existing apps—we inspected the interfaces of ChatGPT, Google Gemini, and Microsoft Copilot, documenting the workflows of privacy-related features with annotated screenshots, and integrating common design patterns such as opting out with a toggle slider into the mockups. During the survey, participants were asked to select the most likely effect of using the *Chat history and training* setting, to determine whether they understood the permanence of data traces in trained models, or expected their data to be removed when they opted out. For the *Delete all chats* and *Delete account* options, we presented participants with the following potential outcomes: losing access to chats or accounts, permanent deletion of chats or accounts, and the permanent removal of patterns the model has learned from chats (see Table 3). We asked participants to select when they thought each outcome would occur given six time points ranging from *Immediately* to *Never*, mirroring the approach used by Schaffner et al. [97] to probe users' mental models of data expiration.

3.2.4 Privacy Awareness and Concern. To address our fourth research question, we explored participants' awareness of and concerns about privacy practices and risks associated with LLM-based CAs. We presented participants with 10 privacy scenarios, and asked whether they had heard of the scenario before completing the survey (prior awareness), whether they thought the scenario was possible (perceptions of plausibility), and how concerned they were about the scenario on a five-point Likert scale, where five was the most concerned. Though privacy scales for directly measuring concern exist (e.g., the IUIPC scale [73]), these are general, and not tailored to specific technologies. Treating privacy awareness and concern as latent variables offered greater flexibility for exploring AI-specific privacy issues, and avoided priming participants with direct references to privacy [14].

We developed our scenarios by reviewing academic literature and case reports on real-world AI privacy incidents. Our scenarios involved a range of data actors (e.g., developers, advertisers, and the government), processing mechanisms (e.g., access and inference), and domains (e.g., health). Many scenarios, such as using chats to train models, represent the existing privacy practices of LLM-based CA apps [83], and others, including memorisation and data extraction have been empirically demonstrated by research studies [20, 81]. Even the more speculative scenarios like targeted health advertising, law enforcement access, and financial profiling are not without precedent: online data is already used in credit scoring [94], and privacy literature has long-illustrated how app data is used to target users with adverts across the web [89]. Indeed, OpenAI has recently expressed an interest in integrating personalised advertising into ChatGPT as the company transitions to a for-profit business model [79]. We integrated these case-studies into short, thematically diverse scenarios which represented real-world situations that users could imagine and respond to. Table 2 presents the scenarios alongside relevant examples.

3.3 Piloting

We conducted two pilot studies to evaluate question comprehensibility, determine the average completion time of the survey, and ensure accessible formatting. The first pilot involved live walk-through sessions with five participants who were asked to think aloud as they completed the questionnaire. They were prompted by the first author to explain how they interpreted questions and decided on their responses. From this, we discovered that some participants were unfamiliar with the concept of model training, motivating us to provide a simple definition at the beginning of the survey. Additionally, our question on information disclosure intentions initially did not prompt participants to consider the relevance of the data items to their task. As a result, many participants found it difficult to rate their willingness to disclose, as they struggled to envision a scenario where such disclosure would be necessary. We revised the question to explicitly include the assumption of task relevance [72]. Finally, we reviewed the privacy scenarios to ensure they were clear to participants with varying technical backgrounds, and that participants' understanding of the scenarios matched the risks and practices being explored, making adjustments to the survey's phrasing as needed. Our second pilot involved distributing the survey to 15 participants to gather feedback on the survey's layout, mobile responsivity, and length. This resulted in minor stylistic changes and improvements to the fidelity of mockup images by increasing element sizing and resolution.

3.4 Quantitative Analysis

We began with a descriptive analysis of participants' responses, using visualisations and summary statistics to provide a broad overview of our findings. We used Friedman's test to detect statistically significant effects of information type on participants' median willingness to disclose data, and of privacy scenarios on their median level of privacy concern. Comparisons were made across the 12 information types described in §3.2.2, and the 10 scenarios described in Table 2 respectively. The information types and scenarios were randomised to account for ordering effects. For detecting main effects, we used Friedman's test as a non-parametric alternative to repeat-measures ANOVA [117], and identified pairwise effects with Holm-adjusted post-hoc tests [40].

We conducted regression analysis to explore the effects of demographics and app usage habits on participants' privacy attitudes and behaviour. Based on results from prior usable privacy studies [11, 26, 71], we identified age, gender, and education level as the most relevant demographic predictors. After checking for multicollinearity, we also included participants' frequency of app usage,

Table 2: Privacy scenarios presented to participants with related examples from academic literature and the press.

Risk/Practice	Scenario	Related Examples
Training	A user's chat history is used to train and improve the company's AI models.	[5, 25, 84]
Memorisation	Information used to train an AI chatbot remains in the model even after the original information is deleted.	[56, 58, 90]
Developer Access	A user's chat history is read by the development company's staff.	[46, 76]
Legal Access	The authorities access someone's chat history and use it as evidence in a criminal investigation.	[75, 99]
Health Inferences	A user experiencing health issues seeks information about their condition from an AI chatbot. Insights drawn from the user's chat logs are sold to third party advertisers, who target them with ads for remedies.	[79, 89]
Financial Inferences	A user interacts with an AI chatbot for budgeting advice. These interactions are analysed, and used to predict the user's financial behaviour. These predictions are sold to credit agencies, impacting the user's credit score.	[94, 107, 109]
Data Extraction Attack	A user with a high level of technical knowledge forces an AI chatbot to output the personal data and chat history of other users.	[20, 81, 93]
Database Leak	A cyberattack on a popular app's database causes user chat histories, names and email addresses to be leaked onto the Internet.	[59, 63, 85]
Insecure Extension	A user downloads a browser extension which automatically optimises their chatbot prompts. Unbeknownst to them, the extension has captured their prompts and leaked them onto the Internet.	[48, 95, 114]
Deception	An AI chatbot manipulates a user into sharing their banking details.	[3, 51]

whether participants used multiple LLM-based CA apps, and their stated purposes for using CAs as predictors. We coded categorical predictors as dummy variables with the most common category used as reference [71]. Reference categories for each variable are indicated in Appendix C.

We ran a binary logistic regression model to predict engagement with each of the eight privacy behaviours. We used the same approach to model whether participants expected data to be unlearned after opting out of model training or deleting data. For information disclosure boundaries and privacy awareness/concerns, we ran binary logistic regression models for each of the 12 information types and 10 privacy scenarios respectively. We collapsed the five Likert items into a two-point binary response to address the sparse distribution of data across the original ordinal scale points. For information disclosure boundaries, we coded a response of 1 (Not at all willing) as 0, and coded the remaining scale points as 1, to conceptually indicate some level of willingness to share the data. For privacy concerns, we modelled high levels of concern by coding the highest categories (4 and 5) as 1. While this approach sacrificed some granularity regarding the strength of participants' attitudes, collapsing low-frequency categories is an effective strategy for handling sparse Likert data and ensuring model convergence when ordinal regression performs unstably [38, 103].

3.5 Qualitative Analysis

We collected a total of 347 responses to our two open questions asking about privacy challenges and potential solutions. Of the provided responses, we excluded 107 from analysis, since they contained a basic 'No' or any phrase which lacked any further content or justification. We analysed the remaining responses qualitatively, using the Dedoose[1] software package for open coding and inter-rater reliability analysis.

The first author coded the responses inductively and generated an initial codebook which was reviewed by the second author. The two authors met to discuss the codebook and made minor initial adjustments. Then, the first and second authors conducted two rounds of independently coding 10% of the dataset, meeting to discuss disagreements and revisions. At this stage, we agreed that certain responses required multiple codes to accurately capture the data, and calculated Cohen's Kappa statistic to quantify inter-rater reliability [74]. Once we reached a sufficiently high Kappa score (>0.7), both authors independently coded the entire dataset, resulting in a final Kappa score of 0.72 (good agreement). The first author re-organised the resulting codes into coherent themes based on topical commonalities and relationships, with the second author providing regular input and feedback. The final codebook we used for our analysis is provided in Appendix D.

3.6 Ethical Considerations

The study was approved under institutional ethics and posed minimal risks to participants. Before starting the survey, participants were provided with an information sheet which summarised the study's purpose, risks, and privacy notice. Ethical risks were minimal: we did not collect any information which could identify users, and all participants were over 18. No vulnerable participants were purposefully recruited for the study. The study did not broach sensitive topics, involve deception, or require participants to use any hardware.

[1] https://www.dedoose.com/

4 Results

4.1 Participants

A total of 211 responses were analysed, with a median completion time of 8.3 minutes. Slightly over half of the respondents were women (52.1%, $n = 110$) and most were aged between 18 and 45 (72.0%, $n = 152$). Over 70% of participants had completed higher education, such as a Bachelor's (50.2%, $n = 106$) or postgraduate degree (20.4%, $n = 43$). All but one participant either currently used ChatGPT, or had done so in the past. Many participants also used Microsoft Copilot (31.3%, $n = 66$) and Google Gemini/Bard (22.7%, $n = 48$). Participants reported diverse use-cases for CAs, including work-related tasks (56.9%, $n = 120$), general information seeking (56.4%, $n = 119$), personal admin (42.2%, $n = 89$), recreational use (49.7%, $n = 105$) and seeking information about personal or health topics (35.5%, $n = 75$).

4.2 RQ1: Privacy Behaviours

Overall, engagement with privacy behaviours was low. On average, participants reported engaging in 1.4 behaviours ($SD = 1.5$) out of the eight presented. As illustrated in Figure 2, the most commonly reported behaviours were deleting/clearing chat histories (31.3%, $n = 66$) and removing sensitive information from prompts (27.0%, $n = 57$). Approximately a quarter of participants had either read the app's privacy policy (23.7%, $n = 50$) and/or learned about how the developers handle personal data from blogs and forums (26.1%, $n = 55$). Of the participants who used an alternative source to learn about privacy, half had not read the privacy policy. Strikingly, only 15 participants (7.1%) had opted out of having their data used to train models. We also explored participants' awareness of privacy features, finding it to be moderate. Approximately half of participants were aware that they could delete their chat history (49.3%, $n = 105$), and slightly fewer were aware that they could export their chat history (46.9%, $n = 99$). Less than a third of participants were aware that they could opt-out of having their data used to train models (28.9%, $n = 61$). However, even among those aware of the opt-out feature, three-quarters still had not used it (41/61).

4.3 RQ2: Information Disclosure Boundaries

Participants were asked to rate their willingness to disclose 12 information types to an LLM-based CA on a scale of 1 to 5. Friedman's test revealed a significant effect of information type on median willingness to disclose ($\chi^2(11) = 1121.0$, $p < 0.01$). The results are visualised in Figure 1, and broadly illustrate that willingness to disclose decreased as the sensitivity of information increased. The darkest regions, representing the greatest percentage of participants, appear at the lowest scale point for banking details, photos, personally identifiable contact details, and credit scores. We note that while personally identifiable contact information was among the data types participants were the least willing to share, an email address is required to create an account with most LLM-based CAs. However, participants rarely used an anonymous email or name to sign up, as we report in §4.2. This may suggest a disconnect between users' disclosure attitudes during chat interactions, and their behaviours during account creation. Participants were willing to share general data such as personality and interests, and some targeted data types such as health and medical history, sexual orientation, and religion or political affiliation. Finally, participants were significantly more willing to share their monthly budget and purchases than their banking details or credit scores, though both data types are finance-related.

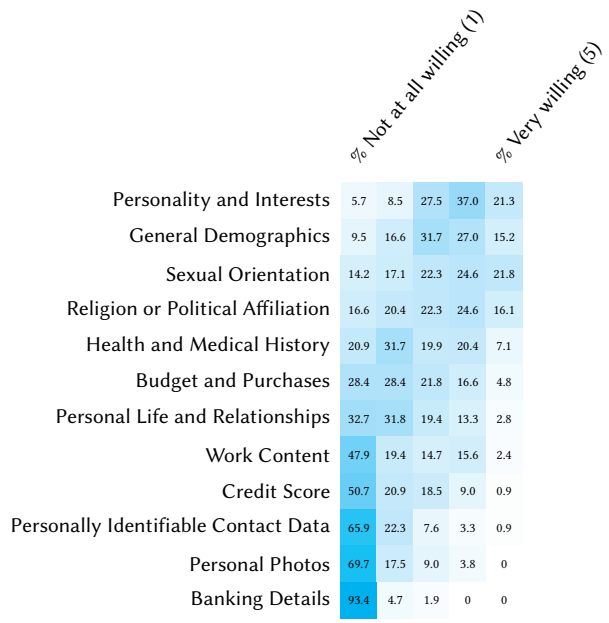


Figure 1: Heatmap of participants' willingness to disclose different types of information with an LLM-based CA.

4.4 RQ3: Comprehension of Privacy Features

Our third research question focused on participants' expectations of three privacy features: 1) Opting out of their data being used to train models; 2) Deleting their chat histories; and 3) Deleting their accounts. First, participants selected the most likely outcome of opting out of model training given three options:

A *My data will not be used to train models in the future, but patterns that the model has learned from my data will* ***remain*** *in the model.*
B *My data will not be used to train models in the future, and patterns that the model has learned from my data will be* ***removed*** *from the model.*
C *My data will continue to be used to train models.*

Over half of participants (54.0%, $n = 114$) selected option ***A***. This response is consistent with the actual outcome of opting out of model training, as implied by the privacy policies and FAQs of developers such as OpenAI: *"While history is disabled, new conversations won't be used to train and improve our models"* [83]. However, over a third (36.0%, $n = 76$) assumed that opting out would not only prevent their data being used to train models in the future, but also remove the influence of past data on the model (option ***B***). Lastly, 10% of participants ($n = 21$) selected option ***C***, believing that the option would have no effect, and that their data would continue to be used to train models.

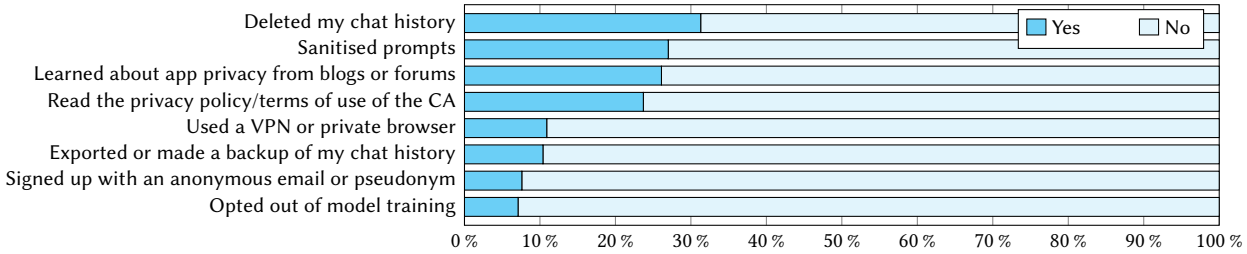


Figure 2: Bar chart illustrating the percentage of participants who engaged in each privacy behaviour.

Participants were also asked about their expectations of chat history and account deletion in terms of when (and if) each outcome would occur. As shown in Table 3, the distribution of responses was similar across the chat history and account deletion scenarios. In both cases, almost all participants assumed that they would immediately lose access to their chat histories and accounts, and that their data would be permanently deleted within a few days. Similarly, almost half of participants expected their data to be removed from the model (i.e., unlearned) either immediately or within a few days.

4.5 RQ4: Awareness and Concerns About Privacy Practices and Risks

To address our fourth research question, we examined participants' awareness of and concerns about ten hypothetical scenarios illustrating various privacy practices and risks associated with LLM-based CAs. Friedman's test revealed a significant effect of privacy scenario on median level of concern ($\chi^2(9) = 423.0$, $p < .001$).

As illustrated in Figure 3, our findings indicate strong risk imagination but lower prior awareness: while participants overwhelmingly thought the scenarios were possible, far fewer had considered them before. For example, while 91.0% ($n = 192$) found developers accessing their chats to be plausible, only 39.8% ($n = 84$) had heard of the practice. Most participants were aware that their data might be used to train and improve models (75.8%, $n = 160$) and that their data could be memorised (51.7%, $n = 109$), but far fewer were aware that developers could access and review their chats. Further, participants were more aware of data being used for targeted health advertising than for financial profiling, and awareness of general data breaches was higher than more specialised scenarios, such as training data extraction or insecure browser extensions. Participants were the least likely to believe that an LLM-based CA could deceive a user into sharing their banking details, and only a fifth of participants were aware of this scenario (21.3%, $n = 45$).

Participants were the most concerned about scenarios with the most adverse outcomes, such as an LLM-based CA deceiving someone into sharing banking details and the inferential use of financial data. Notably, these were also the scenarios participants were the least aware of, indicating reactive concern once the risks were made explicit. By contrast, participants were the least concerned about their chat history being used to train and improve models, with three-quarters (73.9%, $n = 156$) indicating that they were either not at all, or slightly concerned about the practice. Participants were mildly concerned about developers or law enforcement accessing

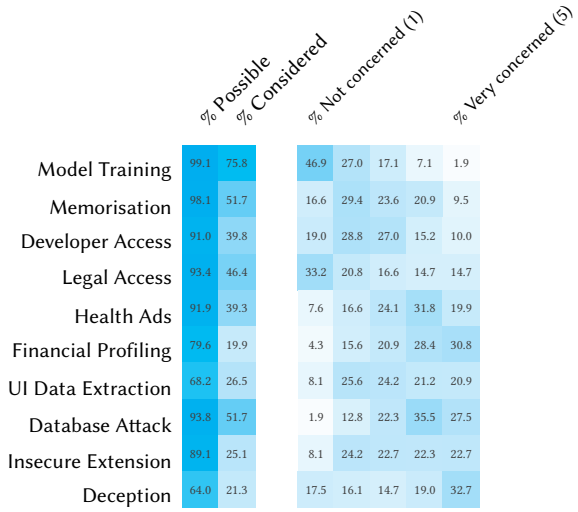


Figure 3: Heatmaps of participants' awareness and concerns about different privacy scenarios.

their chat data, though when data was repurposed for advertising or financial profiling, their concerns increased. This suggests a distinction between data being viewed by reputable data actors versus being reused or analysed for commercial purposes—a more unpredictable use-case which users have limited control over. Of the three attack scenarios, participants were the most aware of and concerned about the database attack, with 63.0% ($n = 133$) rating it as quite or very concerning.

4.6 Effects of Demographics and App Usage

We used regression analysis to determine the effects of demographics and app usage habits on participants' privacy behaviours, information disclosure boundaries, awareness, and concerns. In this section, we summarise statistically significant effects ($p < 0.05$), reporting the odds ratio (OR). We provide the full set of results from all significant models in Appendix E.

4.6.1 Gender. Gender had an effect on privacy behaviours and information disclosure. Women were less likely to have used an anonymous email or pseudonym when signing up for an account with an LLM-based CA ($OR = 0.1$), and had more conservative disclosure boundaries. Specifically, women were less willing to

Table 3: Participants' expectations of when possible outcomes would occur for chat and account deletion.

	% Immediately	% Within days	% Within weeks	% Within months	% Within years	% Never
Chat deletion						
I will no longer be able to access my chat history.	90.0	5.7	1.9	2.4	0	0
My chat history will be permanently deleted.	76.4	11.8	2.8	3.8	0	5.2
Patterns that the model has learned from my data will be removed from the model.	31.8	11.4	3.2	3.8	2.4	47.4
Account deletion						
I will no longer be able to access my account.	87.7	9.0	0.5	1.4	0.5	0.9
My chat history will be permanently deleted.	71.1	11.9	4.8	2.8	2.8	6.6
My account data will be permanently deleted.	67.3	16.6	5.2	3.3	2.4	5.2
Patterns that the model has learned from my data will be removed from the model.	33.1	12.8	2.4	4.3	1.4	46.0

share demographics ($OR = 0.21$), sexual orientation ($OR = 0.25$), banking details ($OR = 0.06$), budget and purchases ($OR = 0.46$), and personal photos ($OR = 0.5$). We found no significant effects of gender on privacy concern or awareness.

4.6.2 Age. Age had an effect on several privacy outcomes. Participants in the 35-44 age range were more likely to opt out of having their data used for model training ($OR = 4.97$) compared to those aged 25-34. However, those in the 25-34 age group were the most likely to sanitise their prompts, surpassing all other groups, including those under 24, who were the least likely to do so ($OR = 0.21$). Participants aged 35 and older were less willing to share their sexual orientation, with the strongest effect observed among those aged 45–54 ($OR = 0.23$). Additionally, those over 55 were less likely to disclose information related to their budget and purchases ($OR = 0.23$). Generally, older participants were less aware of privacy risks: over 45s were less likely to believe that the targeted health advertising scenario was plausible ($OR = 0.05$) and were the least likely to have considered the law enforcement access ($OR = 0.26$) and memorisation scenarios ($OR = 0.3$) before. Younger participants (18 to 24) had a higher level of concern about being deceived by a CA into sharing banking details ($OR = 6.19$).

4.6.3 Level of Education. Participants with a tertiary education were more likely to have read the privacy policy ($OR = 2.94$), read about privacy from another source ($OR = 3.78$), and opted out of having their data used to train models ($OR = 7.81$) compared to those with a Bachelor's degree. Effects on information disclosure were mixed: participants educated up to a high school level were more willing to share their banking details ($OR = 12.22$) and workplace content ($OR = 4.24$) but less willing to share information about their health ($OR = 0.29$). This group also had lower prior awareness of model training ($OR = 0.21$), and a higher level of concern about being deceived into sharing their banking details by a CA ($OR = 4.25$). Participants with a postgraduate level of education were less willing to share photographs ($OR = 0.36$).

4.6.4 Context of App Use. Frequency of use had mixed effects on privacy behaviour. Participants who used LLM-based CAs daily were more likely to have read the privacy policy ($OR = 3.21$). Weekly users were less likely to sign up with a pseudonym or anonymous email ($OR = 0.22$) and opt out of model training than monthly users ($OR = 0.09$). Multi-use predicted higher engagement with several behaviours, including privacy information seeking ($OR = 2.49$), opting out ($OR = 17.8$), prompt sanitisation ($OR = 2.48$), and falsifying sign-up details ($OR = 3.97$). Multi-app users were also less willing to share banking details ($OR = 0.14$), and had a higher prior awareness of model training ($OR = 2.39$).

Finally, participants' self-reported reasons for using LLM-based CAs had significant effects on several privacy outcomes. Firstly, those who used LLM-based CAs for work were more likely to opt-out of model training ($OR = 7.53$) and sanitise their prompts ($OR = 2.36$). Using LLM-based CAs for personal admin purposes predicted a higher willingness to share banking details ($OR = 10.1$) and health information ($OR = 2.68$) and a lower likelihood of believing targeted health advertising was possible ($OR = 0.27$). Interestingly, use of LLM-based CAs for health advice was associated with a much higher likelihood of believing that data could be leaked in a cyberattack. ($OR = 11.5$).

4.7 RQ5: Privacy Challenges and Solutions

To contextualise our quantitative findings, we asked open-ended questions about the privacy challenges participants faced when using LLM-based CAs, and potential solutions. Overall, 151 participants provided a usable response to at least one question. We summarise the key themes from our data in the following section.

4.7.1 Uncertainty, Cynicism, and a Lack of Transparency. Many participants were fearful about AI and lacked confidence in their understanding of app privacy practices ($n = 33$). Practices of concern included storage and retention of chat data ($n = 5$), chats being shared with unauthorised parties ($n = 14$), data linkage and re-identification ($n = 14$), and the generation of inferences or user

profiles from their chats ($n = 7$). They were uncertain about how data was processed behind the scenes, and feared potential misuse: *"There is the uncertainty of never knowing if the owners of these tools will take advantage of my data"* (P123). While some participants attributed their uncertainty to the technical complexity and 'black box' nature of LLMs ($n = 4$), most felt that there was a deliberate lack of transparency regarding how their data was processed and safeguarded ($n = 21$). This lack of transparency made it difficult to formulate an accurate mental model of privacy risks and give informed consent: *"you're not sure what you're even being asked permission for"* (P2).

Building on this, a few participants expressed feelings of cynicism and distrust towards providers ($n = 14$), doubting that tech companies would put user privacy above profits, or considering privacy to be a lost cause: *"I assume I have no protections other than safety in numbers"* (P115). For instance, one participant described AI as the latest thing to degrade user privacy in an imbalanced socio-technical system: *"AI is just the next thing to take more privacy away from people who have very little privacy already"* (P110). These affective dimensions had mixed effects on participants' privacy behaviours. For some, cynicism and distrust motivated vigilance around not sharing sensitive information: *"We are our own first line of defence for protecting our privacy"* (P123). However, other participants responded with resignation, justifying disclosure with the fact that their data was already *"all over the Internet"* (P5), and feeling that nothing could prevent the over-collection of data by LLM developers: *"not using the app doesn't even protect you, you're still hoovered up as a data point"* (P15).

4.7.2 Challenges to Limiting Self-Disclosure. When asked about potential privacy solutions, a large group of participants emphasised individual protective strategies ($n = 67$) which mostly involved some form of limiting disclosure. Participants described three specific strategies: (1) not using CAs for sensitive purposes; (2) redaction, which involved replacing discrete chunks of information (e.g., full names or emails) with placeholders; and (3) abstraction, which involved more nuanced rephrasing or generalisation of prompts to obfuscate details. Although participants' redaction strategies were largely ad hoc and guided by intuition rather than a well-informed privacy threat model, most felt confident that their efforts were adequate to protect their privacy. Many felt that they had nothing to hide, and that the risk of privacy breaches was low as long as they did not share sensitive information: *"As long as you keep anything personal to yourself then they [AI chatbots] can't threaten your privacy"* (P57).

However, some participants acknowledged limitations to their sanitisation strategies, such as being caught off guard by human-like conversations with CAs ($n = 2$) or finding that sanitised prompts produced less relevant outputs ($n = 6$). Optimising the privacy-utility trade-off was challenging in practice: certain tasks like drafting emails required more "targeted" information, and vague language did not provide the results participants needed: *"I work hard to not share personal information, but I feel the responses I get back are too generic"* (P34). A few participants viewed this tension as irreconcilable and believed that privacy loss was a necessary cost: *"The user pays the price of privacy when using chatbots at the gain of conveniently presented data"* (P114). In a similar vein, a few participants feared re-identification from inferences developed through aggregation, analysis, and profiling ($n = 7$). Participants' limited understanding of the model's analytical capabilities made it even more difficult to judge what was safe to share, as inferences were beyond their control and could be drawn from data points they hadn't considered sensitive enough to redact in isolation: *"AI chatbots can make assumptions based on what you ask, it's the unknown and technical side of this that worries me. What information can they gather through simple prompts, we might not even realize we're doing it"* (P136).

4.7.3 Transparency, Education, and Regulatory Solutions. A popular solution mentioned by 47 participants focused on greater transparency, educational initiatives, and accountability from developers. Participants had several knowledge gaps around data storage, including what data was stored and whether data deletion was effective, how data was safeguarded on a technical level, and whether it was shared with third parties or used to train models. However, privacy policies were described as 'deliberately ambiguous' or too long to provide useful answers to their questions, so participants wanted information to be simpler, shorter, and more conspicuous within the app: *"Clearer guidelines about data retention and opt-outs that are more obvious when starting to use the chatbot"* (P62).

Some participants also called for general educational campaigns ($n = 9$) aimed at non-technical users, and practical guidance on what information can be safely shared with LLMs: *"Create a simple Beginners Guide to AI. This can be a short animation or video dispelling common myths about AI etc. There is a lot of mistrust with AI and yet we all use it, we still need more info to educate ourselves"* (P136). However, not all participants trusted developers to be forthcoming with their privacy disclosures. Some called for systemic accountability ($n = 16$), through regulations that prevented chat data from being used for training purposes, enforced the use of encryption, and limited the potential applications of AI: *"new laws need to be implemented to adapt to this advancement in technology"*(P67).

4.7.4 Design Solutions. Finally, 37 participants proposed design suggestions for more usable and transparent privacy features. Suggestions fell into three broad categories: usable privacy settings ($n = 21$), warnings designed to nudge users away from disclosing sensitive information ($n = 11$), and an in-app incognito mode which allowed full use of the app without data being linked to an email address ($n = 5$). Most importantly, participants wanted more discoverable and granular privacy settings. Example suggestions included scheduled auto-deletion for chats, being able to alter model training choices during conversations, and having more control over inferences through being able to access or delete any profiles created about them. Similarly, some participants saw value in interface warnings which reminded them not to share personal information, or automatic tools that alerted users if the prompt they were entering was sensitive. Two participants envisioned a privacy filter which automatically flagged PII and either advised the user on how to make the prompt safer, or automatically redacted the sensitive data: *"there should be warning messages, and the AI should be trained to prevent privacy issues. It should warn users and/or delete personally identifiable information itself"* (P19).

5 Discussion

In this section, we situate our findings within previous literature and discuss implications for policy and design.

5.1 Engagement with Privacy Behaviours and Features

The most common privacy behaviours were deleting chats, sanitising prompts, and seeking privacy information. Even so, less than a third of participants had engaged in these behaviours, and even fewer used VPNs, exported data, or falsified sign-up information, reflecting broader usable privacy trends around low user adoption of formal privacy measures [26, 45, 55, 101]. In particular, only 7% of participants had opted out of their data being used to train models. Previous qualitative studies have identified low feature awareness, negative outcomes (e.g., losing access to chat histories), and a lack of concern about model training as reasons for not opting out [61, 114]. Supporting this, we found that most participants were unaware of the opt-out feature, and perceived model training as the least concerning scenario.

Interestingly, we note that the proportion of participants who self-reported as reading the privacy policy (23.7%) was higher than in several past works [45, 55], with one UK study finding that the click-rate to privacy policies was less than 1% [30]. Firstly, these differences may be explained by our sample, which consisted of regular LLM-based CA users rather than general Internet users. Further, participants might have interpreted the term 'privacy policy' broadly, including materials like FAQs, simplified privacy notices, and app privacy nutrition labels, which they are more likely to have accessed. We also note that self-reported privacy information seeking varied across education levels. Participants educated to a tertiary level were more likely to read privacy policies than participants with a bachelor's degree. While higher educational attainment has been linked to greater privacy literacy [71], previous research has also found that individuals with lower income or education levels may seek out privacy information to compensate for a perceived lack of technical understanding [12, 15].

5.2 Limited Comprehension of Privacy Features

In addition to limited use of available privacy features, data deletion and opting out of model training often did not work as users expected. Our findings build on the work of Zhang et al. [114] by illustrating how inaccurate mental models of inference and training fostered by poor developer transparency, can manifest as misunderstandings of important privacy features. Participants overwhelmingly believed that data would be permanently and immediately erased. However, the standard period for which OpenAI retains deleted chats is 30 days, for instance [83]. Similarly, almost half of participants erroneously believed that either by deleting data or opting out of model training, patterns the model had already learned from their data would be erased (see §2.2). Therefore, while data deletion was the most common privacy behaviour, it was widely misunderstood.

Memorisation, and its implications for data deletion, is not adequately communicated in the interfaces, privacy policies, or FAQs of any popular LLM-based CA, nor is it deeply understood by users [16, 114]. While withholding such information may prevent users from being overwhelmed by technical details, we argue that knowledge of memorisation may influence a user's decision to opt in, share data, or use a particular app [114]. However, beyond informing users and improving opt-out architecture, we also underscore that memorisation and retention must be addressed on a technical level by developers. Current research into machine unlearning is still in the model-centric stage and focuses on remedying the foundational issues discussed in §2.2, rather than exploring how these techniques can be integrated into commercial apps at scale. Designing privacy features that align with users' expectations requires exploring how machine unlearning can be made easier to deploy by app developers (e.g., through reusable code packages) and how they might offer more visible assurances of privacy [113].

5.3 The Challenges of Limiting Self-Disclosure

Participants' willingness to share information with a CA depended on the sensitivity of the information, consistent with findings from several prior studies [23, 70, 113, 114]. Age and gender had strong mediating effects on disclosure willingness, aligning with work by Belen Saglam et al. [10]. In particular, our results favour the interpretation that women have more conservative disclosure boundaries when interacting with LLM-based CAs. While some research suggests women disclose less and are more conscientious about digital hygiene [11, 42], other studies show women may share more data depending on the context and nature of the disclosure [19, 36]. This underscores the need for further research into gender differences in interactions with conversational agents, where communication styles differ from those in social media. In this light, we explored participants' attitudes towards multimodal data in CA interactions, finding that participants—particularly women and those with higher levels of education—were very unwilling to share photographs. LLMs introduce a new paradigm for photo-sharing, distinct from the social media context typically explored in privacy research that focuses on self-presentation and communication trade-offs [8, 33]. To address this emerging use case, future research should investigate users' motivations for sharing images with general-purpose LLM-based CAs, their perceptions of risks, and how text-centric privacy interventions (e.g., [113, 116]) might be adapted for visual data.

In addition to exploring *which* data types users were willing to disclose, we also uncovered insights about *how* disclosures were managed. The redaction and abstraction strategies we identified broadly align with those reported in prior studies [61, 114, 116], and while the standard advice is to refrain from disclosing PII to LLMs [60], our study shows that users have clear disclosure boundaries, and understand the importance of keeping sensitive data private. However, participants often struggled to enforce these boundaries in practice due to cognitive load, uncertainty about what to redact, and concerns about personalised inferences.

5.4 Awareness and Concern About Specific Privacy Practices

We explored participants' awareness of, and concerns about several privacy risks including memorisation, data leakage, access, and personalised inferences. While participants overwhelmingly found the scenarios plausible (indicating an effect of information saliency),

they were much less likely to have considered them before completing the survey. This is despite the fact that some of the scenarios presented, such as data breach incidents, have been widely reported in the press [85]. In particular, we found that participants were more concerned about inferences in health and financial domains, than data being accessed by developers or even law enforcement. This possibly reflects the unpredictability of downstream inferences irrespective of the data actors involved [70, 114]. We also found that participants overwhelmingly did not believe that an LLM-based CA could manipulate a human into divulging their banking details, even though deceptive capabilities of commercial CAs have been documented [3, 49]. Yet, the very same scenario was rated as the most concerning, particularly for younger users and those with lower levels of education, possibly owing to higher financial instability among these user groups [10].

Among all scenarios, participants exhibited the highest prior awareness of, and lowest concerns about model training and memorisation. On the surface, this finding appears to contrast with previous research that suggests users have a low awareness of model training [61, 114]. However, we highlight that while most users have probably considered model training and app improvement in a vague sense, this does not imply an in-depth understanding of how LLMs are trained, or how they might memorise data. We speculate that participants interpreted training in the general sense of using data to improve apps, rather than reflecting on the specific process of training a neural network which introduces technical privacy risks that go beyond typical concerns about sharing analytics with third-parties [111]. Further supporting this intuition, we found that participants were significantly more concerned about the data leakage scenarios, suggesting they may not understand that training can lead to memorisation and data extraction.

5.5 Fear of the Unknown, Distrust, and Limited Choices

Although media coverage and awareness of AI has steadily increased among the general UK public, so has anxiety, with many people describing advancements in AI as 'scary' and feeling that they lack control over how their data is used [1, 34, 35]. Building on the above research, we uncovered that fear and uncertainty about privacy is also felt among regular users of LLM-based CA apps. This indicates that usage of AI does not eliminate privacy concerns, nor does it imply a lack of concern about privacy to begin with. A few participants were distrusting of app developers to use their data responsibly or honour their privacy preferences, and this finding is supported by research indicating that tech companies inspire the lowest levels of trust among the UK public [35]. Therefore, it is vital to address risks both from providers exploiting transparency strategies (e.g., through 'privacy-washing'), and from users dismissing privacy interventions due to distrust and cynicism [1, 35].

Behavioural responses to uncertainty and distrust were diverse. Some participants doubled down on their belief that it was up to individuals to not share risky data with an LLM—an attitude observed in other user studies of smart homes and social media [43, 67]. While this attitude can reflect self-empowerment and confidence, it also risks placing an unfair privacy burden on users [39, 114]. Indeed, other participants responded with cynicism and resigned inaction. We observed that all antecedents of privacy cynicism originally identified by Hoffman et al. [53]—uncertainty, distrust in service providers, and powerlessness—were present to some degree in our findings. Despite provably negative effects on user engagement with privacy [39, 104], cynicism and resignation has yet to be widely explored in the context of generative AI, where data collection is extensive and the processing ever-more complex and opaque. Future research should focus on understanding cynicism among LLM-based app users, and ensuring that these feelings are not misinterpreted as a lack of concern for privacy [114].

5.6 Recommendations for Design and Practice

5.6.1 Strategies for Transparency and Education. Our findings suggest that even regular users of LLM-based CAs experience fear, uncertainty, and misconceptions about privacy. Participants' knowledge gaps commonly centred around data storage, third-party access, data deletion and retention, and how models internally processed their data. While most participants were superficially aware of the concept of model training, our findings suggest that this understanding lacked the accuracy and depth needed for fully informed consent [16, 111]. Educating lay users about how LLMs function requires a careful balance of detail with information utility, and an awareness of how effects may vary across sub-populations of users. At a minimum, users should be made aware of data retention policies, and should understand enough about model training to know that data cannot easily be removed from a trained model.

However, our findings suggest that written materials (e.g., privacy policies) are currently insufficient for addressing these knowledge gaps. Therefore, future work should explore how to embed transparency moments into users' routine app interactions and design for *situated learning* [6], rather than placing the onus on users to read and absorb instructive content. In this way, everyday use of the app can naturally foster correct mental models of LLMs and privacy concepts. For example, data retention can be communicated with alternative post-deletion workflows, such as moving a deleted chat to a separate section where it is marked as scheduled for permanent deletion. During this time, a clear explanation of why the data is being retained can also be shown. Such an approach may help shift users away from a simplistic 'press delete and it's gone' mental model [80], and offer more personalised clarification on whether a user's chats have actually been used in training.

5.6.2 Developing a Design Space for LLM-Specific Privacy Choices. In addition to transparency, users must also be given sufficient choice into how their data is used. While personal autonomy and control over privacy was important to participants, opt-out settings were barely used. Our findings show that the status-quo for privacy settings in popular LLM-based CAs (a single, global toggle) does not offer sufficient flexibility in privacy choices, nor is it enough to address the breadth of privacy concerns that users have. We emphasise the need for novel privacy choice architectures tailored to LLM-based CAs that manage traditional forms of data sharing and use, while also enabling fine-grained control over how chat data is used for model training. For instance, future work could explore how users might set preferences for different types of training—for instance, allowing their data to be used to pre-train models, but opting out of more targeted fine-tuning routines.

In particular, alternative modalities for control which go beyond graphical user interfaces should be explored. In the context of CAs, relying solely on graphical privacy settings forces users to switch modalities to adjust privacy settings, which can increase cognitive load and discourage use of these features [65]. As such, a few studies have explored using chatbots for delivering privacy notice and choice to users, demonstrating that well-timed prompts that offer a specific scope of action (e.g., deleting data at the end of a chat) can increase users' likelihood of engaging with privacy preferences [65], and improve privacy perceptions without sacrificing usability [17]. LLM-based CAs offer even richer opportunities for personalised privacy settings than previous rule-based interventions. More flexible variations of privacy choice architectures can be explored, including different timings and levels of granularity—for example, allowing users to give consent on a chat-by-chat basis or based on specific topics (e.g., prohibiting training on chats relating to health), or timings. Importantly, privacy-preserving defaults and presets should be available, as it is unrealistic to expect that lay users will know which options offer the highest level of privacy [100].

5.6.3 Supporting Participants in Limiting Disclosures. Beyond usable privacy choices, we emphasise that users require guidance on how to interact safely with LLMs while also getting the best from the system. A valuable avenue for future work is context-aware tools that alert users of sensitive disclosures, and suggest improvements or automatically reformulate prompts before they are submitted [114]. Existing approaches include the *Rescriber* tool by Zhou et al. [116] which uses a smaller, distilled LLM to detect, redact, or abstract sensitive prompts and the *Adanonymizer* plugin—a graphical tool that visualises the privacy-utility trade-off for a given prompt as a line graph, allowing users to 'tinker' and iteratively achieve the optimal balance [113]. Both systems provide flexible control options, and illustrate how users can achieve safer disclosures without sacrificing app usability [116]. Importantly, longitudinal evaluations of Rescriber found that users developed a more coherent mental model of appropriate disclosures, and found it easier to balance the privacy-utility trade-off over time. This reaffirms our previous recommendation that practical tools with indirect educational effects may be a more effective strategy than simply providing more information for users to digest.

While insightful, both of the above approaches use a one-size-fits all approach to identifying sensitive data, and match on pre-defined categories of PII. However, users have distinct privacy needs and preferences depending on their tasks, demographics, and levels of experience. Future work should explore how tools can go beyond a static definition of privacy to consider context, defined by the data actors involved, the task being completed, and wider cultural privacy norms [82].

6 Limitations and Future Work

Our sample, while diverse in age and gender, was recruited from Prolific and may differ from the wider population of LLM-based CA users. For instance, most participants had a university education and were users of ChatGPT. This can be explained by the popularity of ChatGPT, and the fact that UK users of AI apps are more likely to be university-educated [35]. However, future work could target apps that are less widely used but attract a more specialised and domain-specific userbase (e.g., healthcare professionals), as these may produce distinct privacy considerations and trust dynamics. While we analysed a broad set of user demographics and attributes, it is likely that participants' technical skills also influenced their privacy attitudes and behaviours—indeed, we found that users of several apps were more engaged with privacy behaviours and had a greater awareness of model training. Therefore, there is value in incorporating a formal measure of technology interest, experience, or comfort in future work. Further, our data was self-reported, and may not reflect participants' true behaviour in practice. Future work could experimentally observe participants' actual interactions with a live prototype, or use data donation methodologies to capture in-the-wild disclosures. In addition, future research could also extend our cross-sectional approach with a longitudinal survey, and explore how changes in user perceptions, knowledge, and experiences with AI can affect privacy attitudes and behaviour over time.

7 Conclusion

In conclusion, we examined the privacy behaviours, awareness, and concerns of LLM-based CA users. Overall, we found that most participants: (1) lacked reliable strategies for safeguarding privacy; (2) struggled to accurately predict the outcomes of privacy features, such as mistakenly believing that data can be unlearned from a trained model; and (3) had a low baseline awareness of many privacy practices and risks, but were concerned about them nonetheless. Our qualitative analysis showed that many participants valued autonomy and responsibility over their own privacy, and embodied this by not sharing sensitive information with CAs. However, this strategy was undermined by challenges in balancing privacy with task effectiveness and navigating implicit disclosures. We identified intervention spaces related to more targeted support for users in correctly understanding how LLMs function, and more usable paradigms for privacy settings with in conversational interfaces.

Acknowledgments

This work was funded by the EPSRC Centre for Doctoral Training in Cybersecurity at University College London (UCL). We would like to thank Dr. Yefim Shulman for advising our analysis of quantitative data, and the anonymous reviewers and editors for their valuable feedback that helped improve our paper. The authors used AI-based tools, including Microsoft Copilot to correct spelling errors and make minor grammatical improvements throughout the paper.

References

[1] Alan Turing Institute, Ada Lovelace Institute. 2022. How do people feel about AI? https://www.turing.ac.uk/how-do-people-feel-about-ai [Accessed: (05-03-2025)].

[2] Moatsum Alawida, Bayan Abu Shawar, Oludare Isaac Abiodun, Abid Mehmood, Abiodun Esther Omolara, and Ahmad K. Al Hwaitat. 2024. Unveiling the Dark Side of ChatGPT: Exploring Cyberattacks and Enhancing User Awareness. *Information* 15, 1 (Jan. 2024), 27. https://doi.org/10.3390/info15010027

[3] Lize Alberts, Ulrik Lyngs, and Max Van Kleek. 2024. Computers as Bad Social Actors: Dark Patterns and Anti-Patterns in Interfaces that Act Socially. *Proceedings of the ACM on Human-Computer Interaction* 8, Cscw1 (April 2024), 1–25. https://doi.org/10.1145/3653693

[4] Shahad Alkamli and Reham Alabduljabbar. 2024. Understanding privacy concerns in ChatGPT: A data-driven approach with LDA topic modeling. *Heliyon* 10, 20 (Oct. 2024), e39087. https://doi.org/10.1016/j.heliyon.2024.e39087

[5] Anthropic. 2025. Notice On Model Training. https://www.anthropic.com/legal/model-training-notice. Accessed: (05-03-2025).

[6] Murat Ataizi. 2012. Situated Learning. In *Encyclopedia of the Sciences of Learning*, Norbert M. Seel (Ed.). Springer US, Boston, MA, 3084–3086. https://doi.org/10.1007/978-1-4419-1428-6_878

[7] Yang Bai, Ge Pei, Jindong Gu, Yong Yang, and Xingjun Ma. 2024. Special Characters Attack: Toward Scalable Training Data Extraction From Large Language Models. https://arxiv.org/abs/2405.05990 _eprint: 2405.05990.

[8] Ardion D. Beldad and Sabrina M. Hegner. 2017. More Photos From Me to Thee: Factors Influencing the Intention to Continue Sharing Personal Photos on an Online Social Networking (OSN) Site among Young Adults in the Netherlands. *International Journal of Human–Computer Interaction* 33, 5 (May 2017), 410–422. https://doi.org/10.1080/10447318.2016.1254890

[9] Rahime Belen Saglam, Jason R. C. Nurse, and Duncan Hodges. 2021. Privacy Concerns in Chatbot Interactions: When to Trust and When to Worry. In *HCI International 2021 - Posters*, Constantine Stephanidis, Margherita Antona, and Stavroula Ntoa (Eds.). Vol. 1420. Springer International Publishing, Cham, 391–399. https://doi.org/10.1007/978-3-030-78642-7_53 Series Title: Communications in Computer and Information Science.

[10] 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), 908245. https://doi.org/10.3389/fcomp.2022.908245

[11] Alex Berke, Badih Ghazi, Enrico Bacis, Pritish Kamath, Ravi Kumar, Robin Lassonde, Pasin Manurangsi, and Umar Syed. 2025. How Unique is Whose Web Browser? The role of demographics in browser fingerprinting among US users. *Proceedings on Privacy Enhancing Technologies* 2025, 1 (Jan. 2025), 720–758. https://doi.org/10.56553/popets-2025-0038

[12] Sophie C. Boerman, Sanne Kruikemeier, and Frederik J. Zuiderveen Borgesius. 2021. Exploring Motivations for Online Privacy Protection Behavior: Insights From Panel Data. *Communication Research* 48, 7 (Oct. 2021), 953–977. https://doi.org/10.1177/0093650218800915

[13] Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine Unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*. 141–159. https://doi.org/10.1109/sp40001.2021.00019

[14] Alex Braunstein, Laura Granka, and Jessica Staddon. 2011. Indirect content privacy surveys: measuring privacy without asking about it. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*. Acm, Pittsburgh Pennsylvania, 1–14. https://doi.org/10.1145/2078827.2078847

[15] Brooke Auxier, Lee Rainie, Monica Anderson, Andrew Perrin, Madhu Kumar, and Erica Turner. 2024. Americans' attitudes and experiences with privacy policies and laws. https://www.pewresearch.org/internet/2019/11/15/americans-attitudes-and-experiences-with-privacy-policies-and-laws/ [Accessed: (18-04-2025)].

[16] Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What Does it Mean for a Language Model to Preserve Privacy?. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2280–2292. https://doi.org/10.1145/3531146.3534642

[17] Birgit Brüggemeier and Philip Lalone. 2022. Perceptions and reactions to conversational privacy initiated by a conversational user interface. *Computer Speech & Language* 71 (Jan. 2022), 101269. https://doi.org/10.1016/j.csl.2021.101269

[18] Tom Buchanan, Carina Paine, Adam N. Joinson, and Ulf-Dietrich Reips. 2007. Development of measures of online privacy concern and protection for use on the Internet. *Journal of the American Society for Information Science and Technology* 58, 2 (Jan. 2007), 157–165. https://doi.org/10.1002/asi.20459

[19] Erin Carbone, George Loewenstein, Irene Scopelliti, and Joachim Vosgerau. 2024. He said, she said: Gender differences in the disclosure of positive and negative information. *Journal of Experimental Social Psychology* 110 (Jan. 2024), 104525. https://doi.org/10.1016/j.jesp.2023.104525

[20] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting Training Data from Large Language Models. http://arxiv.org/abs/2012.07805 arXiv:2012.07805 [cs].

[21] Laurie Carmichael, Sara-Maude Poirier, Constantinos K. Coursaris, Pierre-Majorique Léger, and Sylvain Sénécal. 2022. Users' Information Disclosure Behaviors during Interactions with Chatbots: The Effect of Information Disclosure Nudges. *Applied Sciences* 12, 24 (Dec. 2022), 12660. https://doi.org/10.3390/app122412660

[22] George Chalhoub and Ivan Flechais. 2020. "Alexa, Are You Spying on Me?": Exploring the Effect of User Experience on the Security and Privacy of Smart Speaker Users. In *HCI for Cybersecurity, Privacy and Trust: Second International Conference, HCI-CPT 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings*. Springer-Verlag, Berlin, Heidelberg, 305–325. https://doi.org/10.1007/978-3-030-50309-3_21 event-place: Copenhagen, Denmark.

[23] Paulina Chametka, Sana Maqsood, and Sonia Chiasson. 2023. Security and Privacy Perceptions of Mental Health Chatbots. In *2023 20th Annual International Conference on Privacy, Security and Trust*. Ieee, Copenhagen, Denmark, 1–7. https://doi.org/10.1109/pst58708.2023.10320174

[24] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology* 15, 3 (June 2024), 1–45. https://doi.org/10.1145/3641289

[25] Character AI. 2025. Character.AI Privacy Policy. https://character.ai/privacy. Accessed: (05-03-2025).

[26] Kuanchin Chen and Alan I. Rea Jr. 2004. Protecting Personal Information Online: A Survey of User Privacy Concerns and Control Techniques. *Journal of Computer Information Systems* 44, 4 (2004), 85–92. https://doi.org/10.1080/08874417.2004.11647599 Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/08874417.2004.11647599.

[27] Yufan Chen, Arjun Arunasalam, and Z. Berkay Celik. 2023. Can Large Language Models Provide Security & Privacy Advice? Measuring the Ability of LLMs to Refute Misconceptions. In *Proceedings of the 39th Annual Computer Security Applications Conference (Acsac '23)*. Association for Computing Machinery, New York, NY, USA, 366–378. https://doi.org/10.1145/3627106.3627196 event-place: \textlessconf-loc\textgreater, \textlesscity\textgreaterAustin\textless/city\textgreater, \textlessstate\textgreaterTX\textless/state\textgreater, \textlesscountry\textgreaterUSA\textless/country\textgreater, \textless/conf-loc\textgreater.

[28] Avishek Choudhury and Hamid Shamszare. 2023. Investigating the Impact of User Trust on the Adoption and Use of ChatGPT: Survey Analysis. *Journal of Medical Internet Research* 25 (June 2023), e47184. https://doi.org/10.2196/47184

[29] Jiwon Chung and Hun-yeong Kwon. 2025. Privacy fatigue and its effects on ChatGPT acceptance among undergraduate students: is privacy dead? *Education and Information Technologies* (Jan. 2025). https://doi.org/10.1007/s10639-024-13198-6

[30] Peter Church. 2024. Who reads privacy notices? And why do we have them? https://www.linklaters.com/en/insights/blogs/digilinks/2024/september/uk—who-reads-privacy-notices-and-why-do-we-have-them [Accessed: (18-04-2025)].

[31] Jessica Colnago, Lorrie Faith Cranor, Alessandro Acquisti, and Kate Hazel Jain. 2022. Is it a concern or a preference? an investigation into the ability of privacy scales to capture and distinguish granular privacy constructs. In *Proceedings of the Eighteenth USENIX Conference on Usable Privacy and Security (Soups'22)*. USENIX Association, Usa. event-place: Boston, MA, USA.

[32] Samuel Rhys Cox, Yi-Chieh Lee, and Wei Tsang Ooi. 2023. Comparing How a Chatbot References User Utterances from Previous Chatting Sessions: An Investigation of Users' Privacy Concerns and Perceptions. In *International Conference on Human-Agent Interaction*. Acm, Gothenburg Sweden, 105–114. https://doi.org/10.1145/3623809.3623875

[33] Julia Davies. 2007. Display, Identity and the Everyday: Self-presentation through online image sharing. *Discourse: Studies in the Cultural Politics of Education* 28, 4 (Dec. 2007), 549–564. https://doi.org/10.1080/01596300701625305

[34] Department for Science, Innovation, and Technology. 2023. Public attitudes to data and AI: Tracker survey (Wave 3). https://www.gov.uk/government/publications/public-attitudes-to-data-and-ai-tracker-survey-wave-3/public-attitudes-to-data-and-ai-tracker-survey-wave-3 [Accessed: (05-03-2025)].

[35] Department for Science, Innovation, and Technology. 2024. Public attitudes to data and AI: Tracker survey (Wave 4) report. https://www.gov.uk/government/publications/public-attitudes-to-data-and-ai-tracker-survey-wave-4 [Accessed: (05-03-2025)].

[36] Kathryn Dindia and Mike Allen. 1992. Sex differences in self-disclosure: A meta-analysis. *Psychological Bulletin* 112, 1 (1992), 106–124. https://doi.org/10.1037/0033-2909.112.1.106

[37] Tamara Dinev and Paul Hart. 2006. An Extended Privacy Calculus Model for E-Commerce Transactions. *Information Systems Research* 17, 1 (March 2006), 61–80. https://doi.org/10.1287/isre.1060.0080

[38] Christine DiStefano, Dexin Shi, and Grant B. Morgan. 2021. Collapsing Categories is Often More Advantageous than Modeling Sparse Data: Investigations in the CFA Framework. *Structural Equation Modeling: A Multidisciplinary Journal* 28, 2 (March 2021), 237–249. https://doi.org/10.1080/10705511.2020.1803073

[39] Nora A Draper and Joseph Turow. 2019. The corporate cultivation of digital resignation. *New Media & Society* 21, 8 (Aug. 2019), 1824–1839. https://doi.org/10.1177/1461444819833331

[40] Katie E. Eichstaedt, Keith Kovatch, and David Aaron Maroof. 2013. A less conservative method to adjust for familywise error rate in neuropsychological research: The Holm's sequential Bonferroni procedure. *NeuroRehabilitation* 32, 3 (May 2013), 693–696. https://doi.org/10.3233/nre-130893

[41] Jessica R. Ellis, Natalia S. Dellavalle, Mika K. Hamer, Marlee Akerson, Matt Andazola, Annie A. Moore, Eric G. Campbell, and Matthew DeCamp. 2025. The Halo Effect: Perceptions of Information Privacy Among Healthcare Chatbot Users. *Journal of the American Geriatrics Society* (Feb. 2025), jgs.19393. https://doi.org/10.1111/jgs.19393

[42] Manuela Farinosi and Sakari Taipale. 2018. Who Can See My Stuff? Online Self-Disclosure and Gender Differences on Facebook. *Observatorio (OBS*)* 12, 1

(March 2018). https://doi.org/10.15847/obsOBS12120181129
[43] Casey Fiesler and Blake Hallinan. 2018. "We Are the Product": Public Reactions to Online Data Sharing and Privacy Controversies in the Media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Acm, Montreal QC Canada, 1–13. https://doi.org/10.1145/3173574.3173627
[44] Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. 2024. Membership Inference Attacks against Fine-tuned Large Language Models via Self-prompt Calibration. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=PAWQvrForJ
[45] Sandra Gabriele and Sonia Chiasson. 2020. Understanding Fitness Tracker Users' Security and Privacy Knowledge, Attitudes and Behaviours. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Acm, Honolulu HI USA, 1–12. https://doi.org/10.1145/3313831.3376651 was a priori power analysis conducted to determine sample size?.
[46] Google. 2024. Gemini Apps Privacy Hub. https://support.google.com/gemini/answer/13594961?hl=en-GB#. Accessed: (08-06-2024).
[47] Google. 2024. People + AI Guidebook. https://pair.withgoogle.com/chapter/mental-models/ [Accessed: (08-05-2024)].
[48] Guardio. 2023. "FakeGPT": New Variant of Fake-ChatGPT Chrome Extension Stealing Facebook Ad Accounts with Thousands of Daily Installs. https://labs.guard.io/fakegpt-new-variant-of-fake-chatgpt-chrome-extension-stealing-facebook-ad-accounts-with-4c9996a8f282 [Accessed: (05-03-2025)].
[49] Ece Gumusel, Kyrie Zhixuan Zhou, and Madelyn Rose Sanfilippo. 2024. User Privacy Harms and Risks in Conversational AI: A Proposed Framework. http://arxiv.org/abs/2402.09716 arXiv:2402.09716 [cs].
[50] Maanak Gupta, Charankumar Akiri, Kshitiz Aryal, Eli Parker, and Lopamudra Praharaj. 2023. From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy. *IEEE Access* 11 (2023), 80218–80245. https://doi.org/10.1109/access.2023.3300381
[51] Thilo Hagendorff. 2024. Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences* 121, 24 (June 2024), e2317967121. https://doi.org/10.1073/pnas.2317967121
[52] Franziska Herbert, Florian M. Farke, Marvin Kowalewski, and Markus Dürmuth. 2021. Vision: Developing a Broad Usable Security & Privacy Questionnaire. In *Proceedings of the 2021 European Symposium on Usable Security*. Acm, Karlsruhe Germany, 76–82. https://doi.org/10.1145/3481357.3481526
[53] Christian Pieter Hoffmann, Christoph Lutz, and Giulia Ranzini. 2016. Privacy cynicism: A new approach to the privacy paradox. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 10, 4 (Dec. 2016). https://doi.org/10.5817/cp2016-4-7
[54] Soo Jung Hong. 2025. What drives AI-based risk information-seeking intent? Insufficiency of risk information versus (Un)certainty of AI chatbots. *Computers in Human Behavior* 162 (Jan. 2025), 108460. https://doi.org/10.1016/j.chb.2024.108460
[55] Duha Ibdah, Nada Lachtar, Satya Meenakshi Raparthi, and Anys Bacha. 2021. "Why Should I Read the Privacy Policy, I Just Need the Service": A Study on Attitudes and Perceptions Toward Privacy Policies. *IEEE Access* 9 (2021), 166465–166487. https://doi.org/10.1109/access.2021.3130086
[56] Ico. 2024. How do we ensure individual rights in our AI systems? https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/how-do-we-ensure-individual-rights-in-our-ai-systems/. Accessed: (09-05-2024).
[57] Carolin Ischen, Theo Araujo, Hilde Voorveld, Guda Van Noort, and Edith Smit. 2020. Privacy Concerns in Chatbot Interactions. In *Chatbot Research and Design*, Asbjørn Følstad, Theo Araujo, Symeon Papadopoulos, Effie Lai-Chong Law, Ole-Christoffer Granmo, Ewa Luger, and Petter Bae Brandtzaeg (Eds.). Vol. 11970. Springer International Publishing, Cham, 34–48. https://doi.org/10.1007/978-3-030-39540-7_3 Series Title: Lecture Notes in Computer Science.
[58] Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. 2021. Approximate Data Deletion from Machine Learning Models. https://doi.org/10.48550/arXiv.2002.10077 arXiv:2002.10077 [cs].
[59] Dana Kerr. 2025. DeepSeek hit with 'large-scale' cyber-attack after AI chatbot tops app stores. hhttps://www.theguardian.com/technology/2025/jan/27/deepseek-cyberattack-ai [Accessed: (05-03-2025)].
[60] Hareem Kibriya, Wazir Zada Khan, Ayesha Siddiqa, and Muhammad Khurrum Khan. 2024. Privacy issues in Large Language Models: A survey. *Computers and Electrical Engineering* 120 (Dec. 2024), 109698. https://doi.org/10.1016/j.compeleceng.2024.109698
[61] Angelika Kimbel, Magdalena Glas, and Günther Pernul. 2025. Security and Privacy Perspectives on Using ChatGPT at the Workplace: An Interview Study. In *Human Aspects of Information Security and Assurance*, Nathan Clarke and Steven Furnell (Eds.). Vol. 722. Springer Nature Switzerland, Cham, 184–197. https://doi.org/10.1007/978-3-031-72563-0_13 Series Title: IFIP Advances in Information and Communication Technology.
[62] Spyros Kokolakis. 2017. Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. *Computers & Security* 64 (Jan. 2017), 122–134. https://doi.org/10.1016/j.cose.2015.07.002
[63] Martyn Landi. 2025. ChatGPT maker OpenAI taking claims of data breach 'seriously'. https://www.independent.co.uk/tech/openai-data-breach-chatpt-email-b2694280.html [Accessed: (05-03-2025)].
[64] Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. "I Hear You, I Feel You": Encouraging Deep Self-disclosure through a Chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Acm, Honolulu HI USA, 1–12. https://doi.org/10.1145/3313831.3376175
[65] Anna Leschanowsky, Birgit Popp, and Nils Peters. 2023. Privacy Strategies for Conversational AI and their Influence on Users' Perceptions and Decision-Making. In *Proceedings of the 2023 European Symposium on Usable Security*. Acm, Copenhagen Denmark, 296–311. https://doi.org/10.1145/3617072.3617106
[66] Anna Leschanowsky, Silas Rech, Birgit Popp, and Tom Bäckström. 2024. Evaluating privacy, security, and trust perceptions in conversational AI: A systematic review. *Computers in Human Behavior* 159 (Oct. 2024), 108344. https://doi.org/10.1016/j.chb.2024.108344
[67] Jingjie Li, Kaiwen Sun, Brittany Skye Huff, Anna Marie Bierley, Younghyun Kim, Florian Schaub, and Kassem Fawaz. 2023. "It's up to the Consumer to be Smart": Understanding the Security and Privacy Attitudes of Smart Home Users on Reddit. In *2023 IEEE Symposium on Security and Privacy (SP)*. Ieee, San Francisco, CA, USA, 2850–2866. https://doi.org/10.1109/sp46215.2023.10179344
[68] Yao Li. 2022. Cross-Cultural Privacy Differences. In *Modern Socio-Technical Perspectives on Privacy*, Bart P. Knijnenburg, Xinru Page, Pamela Wisniewski, Heather Richter Lipford, Nicholas Proferes, and Jennifer Romano (Eds.). Springer International Publishing, Cham, 267–292. https://doi.org/10.1007/978-3-030-82786-1_12
[69] Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, Kush R. Varshney, Mohit Bansal, Sanmi Koyejo, and Yang Liu. 2025. Rethinking machine unlearning for large language models. *Nature Machine Intelligence* 7, 2 (Feb. 2025), 181–194. https://doi.org/10.1038/s42256-025-00985-0
[70] Zhihuang Liu, Ling Hu, Tongqing Zhou, Yonghao Tang, and Zhiping Cai. 2025. Prevalence Overshadows Concerns? Understanding Chinese Users' Privacy Awareness and Expectations Towards LLM-based Healthcare Consultation. In *2025 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, Los Alamitos, CA, USA, 92–92. https://doi.org/10.1109/sp61157.2025.00092 Issn: 2375-1207.
[71] Byron Lowens, Sean Scarnecchia, Jane Im, Tanisha Afnan, Annie Chen, Yixin Zou, and Florian Schaub. 2025. Misalignments and Demographic Differences in Expected and Actual Privacy Settings on Facebook. *Proceedings on Privacy Enhancing Technologies* 2025, 1 (Jan. 2025), 456–471. https://doi.org/10.56553/popets-2025-0025
[72] Miguel Malheiros, Sacha Brostoff, Charlene Jennett, and Angela Sasse. 2013. Would You Sell Your Mother's Data? Personal Data Disclosure in a Simulated Credit Card Application. *The Economics of Information Security and Privacy* (Oct. 2013). https://doi.org/10.1007/978-3-642-39498-0_11 Isbn: 978-3-642-39497-3.
[73] Naresh K. Malhotra, Sung S. Kim, and James Agarwal. 2004. Internet Users' Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model. *Information Systems Research* 15, 4 (Dec. 2004), 336–355. https://doi.org/10.1287/isre.1040.0032
[74] Marry L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica* (2012), 276–282. https://doi.org/10.11613/bm.2012.031
[75] Meta. 2025. Information for law enforcement authorities. https://about.meta.com/actions/safety/audiences/law/guideliness [Accessed: (05-03-2025)].
[76] Microsoft Azure. 2025. Abuse Monitoring. https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/abuse-monitoring?source=recommendations [Accessed: (05-03-2025)].
[77] Niloofar Mireshghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. 2024. Trust No Bot: Discovering Personal Disclosures in Human-LLM Conversations in the Wild. https://doi.org/10.48550/arXiv.2407.11438 arXiv:2407.11438 [cs].
[78] David L. Mothersbaugh, William K. Foxx, Sharon E. Beatty, and Sijun Wang. 2012. Disclosure Antecedents in an Online Service Context: The Role of Sensitivity of Information. *Journal of Service Research* 15, 1 (Feb. 2012), 76–98. https://doi.org/10.1177/1094670511424924
[79] Madhumita Murgia, Cristina Criddle, and Hammone. 2024. OpenAI explores advertising as it steps up revenue drive. https://www.ft.com/content/9350d075-1658-4d3c-8bc9-b9b3dfc29b26 [Accessed: (05-03-2025)].
[80] Ambar Murillo, Andreas Kramm, Sebastian Schnorf, and Alexander De Luca. 2018. "If I press delete, it's gone" - User Understanding of Online Data Deletion and Expiration. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*. USENIX Association, Baltimore, MD, 329–339. https://www.usenix.org/conference/soups2018/presentation/murillo
[81] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian

Tramèr, and Katherine Lee. 2023. Scalable Extraction of Training Data from (Production) Language Models. http://arxiv.org/abs/2311.17035 arXiv:2311.17035 [cs].

[82] Helen Nissenbaum. 2004. Privacy as Contextual Integrity. *Washington Law Review* 79, 1 (2004).

[83] OpenAI. 2024. Europe privacy policy. https://openai.com/en-GB/policies/eu-privacy-policy/. Accessed: (08-06-2024).

[84] OpenAI. 2024. How your data is used to improve model performance. https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance

[85] OpenAI. 2024. March 20 ChatGPT outage: Here's what happened. https://openai.com/blog/march-20-chatgpt-outage. Accessed: (08-05-2024).

[86] Eva Orszaghova and Grant Blank. 2024. Does the type of privacy-protective behaviour matter? An analysis of online privacy protective action and motivation. *Information, Communication & Society* 27, 14 (Oct. 2024), 2530–2547. https://doi.org/10.1080/1369118x.2024.2334906

[87] Kentrell Owens, Johanna Gunawan, David Choffnes, Pardis Emami-Naeini, Tadayoshi Kohno, and Franziska Roesner. 2022. Exploring Deceptive Design Patterns in Voice Interfaces. In *Proceedings of the 2022 European Symposium on Usable Security (EuroUSEC '22)*. Association for Computing Machinery, New York, NY, USA, 64–78. https://doi.org/10.1145/3549015.3554213 event-place: Karlsruhe, Germany.

[88] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael P. Wellman. 2018. SoK: Security and Privacy in Machine Learning. In *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*. Ieee, London, 399–414. https://doi.org/10.1109/EuroSP.2018.00035

[89] Lisa Parker, Vanessa Halter, Tanya Karliychuk, and Quinn Grundy. 2019. How private is your mental health app data? An empirical study of mental health app privacy policies and practices. *International Journal of Law and Psychiatry* 64 (May 2019), 198–204. https://doi.org/10.1016/j.ijlp.2019.04.002

[90] Vaidehi Patil, Peter Hase, and Mohit Bansal. 2023. Can Sensitive Information Be Deleted From LLMs? Objectives for Defending Against Extraction Attacks. https://doi.org/10.48550/arXiv.2309.17410 arXiv:2309.17410 [cs].

[91] Md. Abdur Rahman. 2023. A Survey on Security and Privacy of Multimodal LLMs - Connected Healthcare Perspective. In *2023 IEEE Globecom Workshops (GC Wkshps)*. Ieee, Kuala Lumpur, Malaysia, 1807–1812. https://doi.org/10.1109/GCWkshps58843.2023.10465035

[92] Kopo Marvin Ramokapane, Awais Rashid, and Jose Miguel Such. 2017. "I feel stupid I can't delete...": A Study of Users' Cloud Deletion Practices and Coping Strategies. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*. USENIX Association, Santa Clara, CA, 241–256. https://www.usenix.org/conference/soups2017/technical-sessions/presentation/ramokapane

[93] Siladitya Ray. 2024. Samsung Bans ChatGPT Among Employees After Sensitive Code Leak. https://www.forbes.com/sites/siladityaray/2023/05/02/samsung-bans-chatgpt-and-other-chatbots-for-employees-after-sensitive-code-leak/ [Accessed: (08-05-2024)].

[94] RiskSeal. 2025. How Social Media Profiling Enhances Credit Risk Management. https://riskseal.io/blog/how-social-media-profiling-enhances-credit-risk-management#toc-use-of-social-media-activity-for-credit-assessments [Accessed: (05-03-2025)].

[95] Guy Rosen. 2023. Meta's Q1 2023 Security Reports: Protecting People and Businesses. http://about.fb.com/news/2023/05/metas-q1-2023-security-reports/ [Accessed: (05-03-2025)].

[96] Ali Satvaty, Suzan Verberne, and Fatih Turkmen. 2024. Undesirable Memorization in Large Language Models: A Survey. https://doi.org/10.48550/arXiv.2410.02650 arXiv:2410.02650 [cs].

[97] Brennan Schaffner, Neha A. Lingareddy, and Marshini Chetty. 2022. Understanding Account Deletion and Relevant Dark Patterns on Social Media. *Proceedings of the ACM on Human-Computer Interaction* 6, Cscw2 (Nov. 2022), 1–43. https://doi.org/10.1145/3555142

[98] Scott Schanke, Gordon Burtch, and Gautam Ray. 2021. Estimating the Impact of "Humanizing" Customer Service Chatbots. *Information Systems Research* 32, 3 (Sept. 2021), 736–751. https://doi.org/10.1287/isre.2021.1015

[99] Jon Schuppe. 2022. Police sweep Google searches to find suspects. The tactic is facing its first legal challenge. https://www.nbcnews.com/news/us-news/police-google-reverse-keyword-searches-rcna35749. Accessed: (05-03-2025).

[100] Johanneke Siljee. 2015. Privacy transparency patterns. In *Proceedings of the 20th European Conference on Pattern Languages of Programs (EuroPLoP '15)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2855321.2855374 event-place: Kaufbeuren, Germany.

[101] Nissy Sombatruang, Tan Omiya, Daisuke Miyamoto, M. Angela Sasse, Youki Kadobayashi, and Michelle Baddeley. 2020. Attributes Affecting User Decision to Adopt a Virtual Private Network (VPN) App. In *Information and Communications Security*, Weizhi Meng, Dieter Gollmann, Christian D. Jensen, and Jianying Zhou (Eds.). Springer International Publishing, Cham, 223–242.

[102] Anna Stock, Stephan Schlögl, and Aleksander Groth. 2023. Tell Me, What Are You Most Afraid Of? Exploring the Effects of Agent Representation on Information Disclosure in Human-Chatbot Interaction. In *Artificial Intelligence in HCI*, Helmut Degen and Stavroula Ntoa (Eds.). Vol. 14051. Springer Nature Switzerland, Cham, 179–191. https://doi.org/10.1007/978-3-031-35894-4_13 Series Title: Lecture Notes in Computer Science.

[103] Chia-Lin Tsai, Stefanie Wind, and Samantha Estrada. 2025. Exploring the Effects of Collapsing Rating Scale Categories in Polytomous Item Response Theory Analyses: An Illustration and Simulation Study. *Measurement: Interdisciplinary Research and Perspectives* 23, 1 (Jan. 2025), 66–89. https://doi.org/10.1080/15366367.2023.2288791

[104] Iris Van Ooijen, Claire M. Segijn, and Suzanna J. Opree. 2024. Privacy Cynicism and its Role in Privacy Decision-Making. *Communication Research* 51, 2 (March 2024), 146–177. https://doi.org/10.1177/00936502211060984

[105] Kaicheng Wang. 2024. From ELIZA to ChatGPT: A brief history of chatbots and their evolution. *Applied and Computational Engineering* 39, 1 (Feb. 2024), 57–62. https://doi.org/10.54254/2755-2721/39/20230579

[106] Lijin Wang, Jingjing Wang, Jie Wan, Lin Long, Ziqi Yang, and Zhan Qin. 2024. Property Existence Inference against Generative Models. In *33rd USENIX Security Symposium (USENIX Security 24)*. USENIX Association, Philadelphia, PA, 2423–2440. https://www.usenix.org/conference/usenixsecurity24/presentation/wang-lijin

[107] Yanhao Wei, Pinar Yildirim, Christophe Van den Bulte, and Chrysanthos Dellarocas. 2016. Credit Scoring with Social Network Data. *Marketing Science* 35, 2 (2016), 234–258. http://www.jstor.org/stable/44012148 Publisher: INFORMS.

[108] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. Acm, Seoul Republic of Korea, 214–229. https://doi.org/10.1145/3531146.3533088

[109] Wilhelmina Afua Addy, Adeola Olusola Ajayi-Nifise, Binaebi Gloria Bello, Sunday Tubokirifuruar Tula, Olubusola Odeyemi, and Titilola Falaiye. 2024. AI in credit scoring: A comprehensive review of models and predictive analytics. *Global Journal of Engineering and Technology Advances* 18, 2 (Feb. 2024), 118–129. https://doi.org/10.30574/gjeta.2024.18.2.0029

[110] Meredydd Williams and Jason R. C. Nurse. 2016. Optional Data Disclosure and the Online Privacy Paradox: A UK Perspective. In *Human Aspects of Information Security, Privacy, and Trust*, Theo Tryfonas (Ed.). Vol. 9750. Springer International Publishing, Cham, 186–197. https://doi.org/10.1007/978-3-319-39381-0_17 Series Title: Lecture Notes in Computer Science.

[111] Amy Winograd. 2023. Loose-Lipped Large Language Models Spill Your Secrets: The Privacy Implications of Large Language Models. *Harvard Journal of Law & Technology* 36, 2 (2023). https://jolt.law.harvard.edu/assets/articlePDFs/v36/Winograd-Loose-Lipped-LLMs.pdf

[112] Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S. Yu. 2023. Machine Unlearning: A Survey. *ACM Comput. Surv.* 56, 1 (Aug. 2023). https://doi.org/10.1145/3603620 Place: New York, NY, USA Publisher: Association for Computing Machinery.

[113] Shuning Zhang, Xin Yi, Haobin Xing, Lyumanshan Ye, Yongquan Hu, and Hewu Li. 2025. Adanonymizer: Interactively Navigating and Balancing the Duality of Privacy and Output Performance in Human-LLM Interaction. https://doi.org/10.48550/arXiv.2410.15044 arXiv:2410.15044 [cs].

[114] 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 *Proceedings of the CHI Conference on Human Factors in Computing Systems (Chi '24)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3613904.3642385

[115] Junhao Zhou, Yufei Chen, Chao Shen, and Yang Zhang. 2021. Property Inference Attacks Against GANs. https://doi.org/10.48550/arXiv.2111.07608 arXiv:2111.07608 [cs].

[116] Jijie Zhou, Eryue Xu, Yaoyao Wu, and Tianshi Li. 2025. Rescriber: Smaller-LLM-Powered User-Led Data Minimization for LLM-Based Chatbots. https://doi.org/10.1145/3706598.3713701 arXiv:2410.11876 [cs].

[117] Donald W. Zimmerman and Bruno D. Zumbo. 1993. Relative Power of the Wilcoxon Test, the Friedman Test, and Repeated-Measures ANOVA on Ranks. *The Journal of Experimental Education* 62, 1 (July 1993), 75–86. https://doi.org/10.1080/00220973.1993.9943832

[118] V.W. Zue and J.R. Glass. 2000. Conversational interfaces: advances and challenges. *Proc. IEEE* 88, 8 (2000), 1166–1180. https://doi.org/10.1109/5.880078

A Survey Questionnaire

A.1 Demographics

Q1. How old are you?

- ❍ 18-24
- ❍ 25-34
- ❍ 35-44
- ❍ 45-54
- ❍ 55+

Q2. What is your gender?

- ❍ Woman
- ❍ Man
- ❍ Non-binary
- ❍ Prefer to self-describe

Q3. What is your ethnic group?

- ❍ White
- ❍ Black
- ❍ Asian
- ❍ Arab/North African
- ❍ Mixed, please specify
- ❍ Prefer to self-describe

Q4. What is your highest level of education?

- ❍ High School
- ❍ A-Level or equivalent
- ❍ Bachelor's degree
- ❍ Master's degree
- ❍ Doctorate/PhD

Q5. Please describe your job/profession. If you are a student, please specify the subject that you are currently studying.

A.2 Chatbot Usage

* **Q6.** Which of the following AI chatbots have you interacted with?

- ❑ ChatGPT
- ❑ Google Gemini/Bard
- ❑ Microsoft Copilot/Bing
- ❑ Github Copilot
- ❑ Claude
- ❑ Other, please specify

* **Q7.** How often do you use the above chatbots?

- ❍ Daily
- ❍ Weekly
- ❍ Monthly

* **Q8.** How long have you been a user of AI chatbots?

- ❍ Less than 1 month
- ❍ 1-3 months
- ❍ 3-6 months
- ❍ 6-12 months
- ❍ More than 12 months

* **Q9.** How do you access AI chatbots?

- ❍ Desktop/web app
- ❍ Mobile app
- ❍ Plugin/Browser extension
- ❍ Application Programming Interface (API)
- ❍ Other, please specify

* **Q10.** What do you typically use AI chatbots for?

- ❑ Work or job-related tasks
- ❑ Personal admin
- ❑ Information or advice about general topics
- ❑ Information or advice about personal topics
- ❑ Creative purposes
- ❑ Other, please specify

A.3 Privacy Behaviours

* **Q12.** During my time using an AI chatbot, I have...

- ❑ Read the privacy policy/terms of use of the chatbot
- ❑ Read about how AI chatbots handle user data from another source, such as blogs, articles, or forums.
- ❑ Removed personal information from prompts.
- ❑ Deleted/cleared my chat history.
- ❑ Exported or made a backup of my chat history.
- ❑ Opted out of my chat history being used to train models.
- ❑ Created an anonymous email and/or pseudonym to use an AI chatbot so that my chats are not linked to me.
- ❑ Used a VPN or private browser while using an AI chatbot to make it harder to track me.

* **Q13.** Before completing this survey, were you aware of the following AI chatbot features?

- ❑ Deleting chat history
- ❑ Deleting account
- ❑ Exporting chat history
- ❑ Opting out of chat history being used to train models.

A.4 Information Disclosure

The following questions are about an app called ConvoGenie, a fictional AI chatbot. Please assume that you use ConvoGenie regularly for general tasks and inquiries. The main interface is shown below.

* **Q14** Please indicate how willing you would be to share the following pieces of information with ConvoGenie during a conversation where that information is relevant.)

Participants rated their willingness to share the following information types with 1=Not at all willing, 2=Slightly willing, 3=Moderately willing, 4=Quite willing, 5=Very willing.

- Personally identifiable contact information such as your name, phone number, or email address
- Your personality and interests, such as your favourite books, movies, or hobbies
- Demographic information such as your age, race, or gender
- Your religion or political affiliation
- Your sexual orientation
- Information about your personal life and relationships
- Information about your health, fitness, or medical history
- Pictures of you or people you know
- Your monthly budget and purchases
- Your banking information (e.g., card details)
- Your credit score
- Work-related content such as reports, documents, or code

A.5 Expectations of Privacy Features

The next screenshot depicts the settings for your ConvoGenie account. The following questions will ask about your expectations of how these settings work.

* **Q15.** When I disable *Chat history and training* I expect...
  - My data will not be used to train models in the future, and patterns that the model has learned from my data will be removed from the model.
  - My data will not be used to train models in the future, but patterns that the model has learned from my data will remain in the model
  - My data will continue to be used to train models

* **Q16.** When I use the *Delete all chats* option, I expect...
  - I will no longer be able to access my chat history.
  - My chat history will be permanently deleted.
  - All patterns that the model has learned from my chat history will be All patterns that the model has learned from my chat history will be **removed** from the model.

* **Q17.** When I use the *Delete account* option, I expect...
  - I will no longer be able to access or log into my account.
  - My account information (email, phone number, banking details) will be permanently deleted.
  - My chat history will be permanently deleted.
  - All patterns that the model has learned from my chat history/data will be All patterns that the model has learned from my chat history will be **removed** from the model.

For each outcome, participants selected one from the following: [Immediately, Within a few days, Within a few months, Within a few years, This will never happen.]

A.6 Awareness and Concern about Privacy Risks

For each scenario (see Table 2), the following questions were asked:

* **Q18.** Have you ever thought about this scenario (or a similar one) before completing this survey?
  - ❍ Yes
  - ❍ No

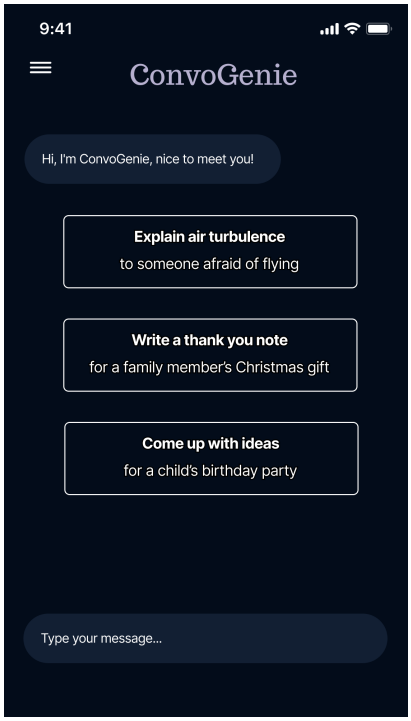
* **Q19.** How concerned are you about this scenario?
  - ❍ Not at all concerned
  - ❍ A little concerned
  - ❍ Moderately concerned
  - ❍ Quite concerned
  - ❍ Very concerned

* **Q20.** Do you think this scenario is possible?
  - ❍ Yes
  - ❍ No

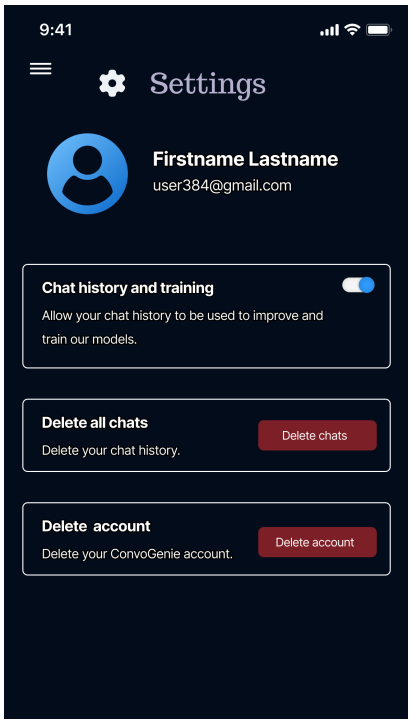
A.7 Open Questions

* **Q21.** Do you find anything challenging about protecting your privacy when using AI chatbots? If so, what?

* **Q22.** Is there anything that you feel would help you to protect your privacy when using AI chatbots?



(a) Mocked up homepage



(b) Mocked up settings page

Figure 4: Mocked up interface for the fictional LLM-based CA *ConvoGenie*

B Development of Disclosure Items

We developed our information types by adapting prior studies of self-disclosure in the context of chatbot interactions [10, 21, 23], e-commerce transactions [78], and online credit card applications [72]. After collating data items and removing duplicates/items that did not make sense to share with a CA, we were left with 42 items, which we organised into eight categories:

- **PII and Contact Details:** Full Name, Home Address, Phone Number, Email, GPS Location
- **Basic Demographics:** Gender, DOB, Education Level, Occupation, Racial or Ethnic Origin, Nationality
- **Special Demographics:** Political Affiliation, Religion, Sexual Orientation, Trade Union Membership
- **Health and Medical History:** Substance Usage, Medical History and Conditions, HIV Status, Sleep Issues, Concentration Issues, Emotional Difficulties/Mental Health
- **Personal Life and Relationships:** Mother's Maiden Name, Family Structure, Marital Status, PII of Friends, Relationship History, Relationship Issues, Sex Life, Criminal Record
- **Basic Financial:** Income Level, Value of Home, Employment Status, Budget
- **Sensitive Financial:** Credit Card Number, Credit Score, Bank Account Credentials, Bills and Payment Histories, Insurance Claims
- **Personality and Interests:** Internet Usage Habits, Leisure and Hobbies, Favourite Products/Brands, Media Consumption Habits

We amalgamated the above categories, and split religious/political affiliation and sexual orientation into two categories. We did the same for banking information and credit scores. We added two more data items to reflect the multimodality of LLMs: Photos and Documents (e.g., PDFs). This resulted in twelve information types overall, described in full in A.4.

C Participant Demographics

Table 4: Summary of participants' demographics and app usage habits. Items marked with an asterisk were used as the reference category for regression modeling. The *CAs used* and *Reasons for using CAs* questions were multi-select, and so each option was treated as a binary variable.

	#	%
Gender		
Woman	110	52.1
Man	98	46.4
Non-binary	3	1.3
Age		
18-24	24	11.4
25-34 *	68	32.0
35-44	60	28.3
45-54	33	15.6
55+	26	12.3
Highest level of education		
High school	20	9.5
A-Level or tertiary	42	19.9
Bachelor's degree *	106	50.2
Postgraduate degree	43	20.4
Frequency of CA use		
Daily	31	14.7
Weekly	83	39.3
Monthly *	97	46.0
CAs Used		
ChatGPT	209	99.1
Microsoft Copilot	66	31.3
Google Gemini	48	22.7
Other	30	14.2
Reasons for using CAs		
Work	120	56.9
Administrative tasks (e.g., finances)	89	42.2
Health and personal advice	75	35.5
General information seeking and advice	119	56.4
Recreational	105	49.7
Other	12	5.7

D Qualitative Analysis Codebook

Table 5: Final codebook for qualitative survey analysis.

	Code name	Description	Example data extract
Attitudes and Challenges	Concern and Uncertainty	An expression of uncertainty or anxiety towards privacy practices, or an self-perceived lack of technical knowledge and privacy self-efficacy.	*"I think there is still some unknown about where the data is stored and who has access to it." (P85)*
	Lack of Transparency	Perceptions that developers/companies do not provide enough open or usable information about how user data is stored, processed, or shared.	*"It's challenging to understand what the ai can do, terms and conditions can be deliberately long and ambiguous."* (P87)
	Cynicism and Distrust	The belief that privacy is unattainable, and a lack of trust in developers' assurances to safeguard data.	*"I'm not sure that any tech company's assurances have ever proven to be worth the paper they're presumably not written on."* (P115)
Personal Strategies	Limiting Disclosures	Making a conscious effort to limit the sensitive data shard with a CA, through redacting prompts or not using CAs for personal reasons.	*"I have only used chatbots for quite trivial tasks so have not worried excessively about privacy."* (P96)
	Using PETs	Using specific privacy enhancing technologies when interacting with LLM-based CAs such as a VPN or private browser.	*"VPN and incognito mode would help."* (P33)
	Anonymous Accounts	Signing up for an account with a pseudonym or separate anonymous email to avoid linking chats to personal identity.	*"Perhaps make a fake account so you don't even use your own email."* (P76)
Governance Solutions	Transparency and Education	Calls for increased transparency from developers surrounding privacy practices, and wider educational campaigns for raising public awareness about AI.	*"Clearer information about where your information is stored and who can access it."* (P37)
	Regulations	A desire for targeted regulations compelling companies to safeguard user data and limit its secondary use.	*"Reassurance from government bodies about legal rights and polices in place to help protect users."* (P53)
Design-based Solutions	Warnings and filters	Warnings on the interface to remind users not to share PII, and automatic removal of PII in prompts.	*"There should be warning messages, and the ai should be trained to prevent privacy issues... ie. it should warn users, and/or delete personally identifiable information itself."* (P19)
	Usable privacy controls	More transparent and user-friendly options for deleting data or opting out, e.g., in-conversation privacy controls and scheduled autodeletion.	*"In each chat, you choose how in that exact scenario your data is used."* (P87)

E Regression Analysis

E.1 Privacy Behaviour

Table 6: Results for significant binary logistic regression models predicting engagement with privacy behaviours. As indicated by column names, the behaviours with significant demographic predictors were reading the privacy policy, reading an alternative source to learn about privacy, opting out of model training, sanitising prompts, and falsifying signup data.

	Read Privacy Policy $LL = -100.9$		**Read Alt. Source** $LL = -108.8$		**Opt-Out** $LL = -38.6$		**Prompt Sanitisation** $LL = -108.02$		**Data Falsification** $LL = -43.7$	
	OR	*p*	*OR*	*p*	*OR*	*p*	*OR*	*p*	*OR*	*p*
Woman	0.57	0.13	0.54	0.08	0.78	0.70	1.10	0.78	0.10	**0.005**
Age (18-24)	0.31	0.12	0.65	0.49	0.46	0.53	0.21	**0.017**	0.24	0.23
Age (35 - 44)	0.64	0.35	0.68	0.38	4.97	**0.045**	0.34	**0.015**	0.46	0.29
Age (45 - 54)	2.25	0.11	0.78	0.64	0.26	0.32	0.30	**0.027**	0.17	0.08
Age (55+)	0.40	0.19	0.53	0.32	0.84	0.90	0.41	0.16	0.26	0.27
Edu. (HS)	3.25	0.06	1.56	0.49	0.83	0.89	1.07	0.92	0.33	0.40
Edu. (A-Level)	2.94	**0.023**	3.78	**0.004**	7.81	**0.014**	2.10	0.11	1.60	0.54
Edu. (Postgraduate)	0.52	0.22	1.22	0.66	1.65	0.57	1.45	0.40	0.87	0.87
Freq. of Use (Daily)	3.21	**0.048**	1.42	0.51	0.41	0.30	1.07	0.89	0.49	0.45
Freq. of Use (Weekly)	1.45	0.38	0.90	0.79	0.09	**0.015**	0.62	0.24	0.22	**0.047**
Multi-user	1.28	0.53	2.49	**0.014**	17.79	**0.002**	2.48	**0.014**	3.97	**0.045**
Work use	1.43	0.40	1.61	0.24	7.53	**0.029**	2.36	**0.033**	1.01	0.99
Personal use	1.24	0.58	1.24	0.55	2.10	0.25	1.59	0.19	1.84	0.34
Health use	1.44	0.33	1.07	0.86	0.62	0.49	1.31	0.48	1.53	0.50

E.2 Information Disclosure

Table 7: Results for significant binary logistic regression models predicting willingness to disclose information. As indicated by column names, the data types with significant demographic predictors were demographics, sexual orientation, banking details, budget and purchases, health information, personal photos, and workplace material.

	Demographics $LL = -50.55$		**S. Orientation** $LL = -73.5$		**Banking** $LL = -34.2$		**Budget** $LL = -112.2$		**Health** $LL = -95.6$		**Photos** $LL = -117.6$		**Work** $LL = -133.8$	
	OR	*p*	*OR*	*p*	*OR*	*p*	*OR*	*p*	*OR*	*p*	*OR*	*p*	*OR*	*p*
Woman	0.21	**0.017**	0.25	**0.005**	0.06	**0.003**	0.46	**0.037**	0.58	0.18	0.50	**0.037**	0.86	0.62
Age (18-24)	0.44	0.34	0.56	0.47	1.82	0.59	2.92	0.19	1.97	0.36	0.47	0.26	0.58	0.30
Age (35 - 44)	0.76	0.72	0.29	**0.048**	2.89	0.25	0.62	0.28	1.22	0.69	1.05	0.90	0.86	0.70
Age (45 - 54)	0.48	0.35	0.23	**0.033**	2.27	0.46	0.49	0.15	0.86	0.78	2.02	0.15	0.93	0.87
Age (55+)	0.23	0.11	0.32	0.15	1.38	0.83	0.23	**0.009**	1.15	0.82	1.11	0.85	0.95	0.92
Edu. (HS)	2.81	1.00	2.01	0.43	12.22	**0.026**	1.35	0.62	0.29	**0.045**	1.87	0.26	4.24	**0.018**
Edu. (A-Level)	9.04	1.00	1.94	0.30	0.65	0.67	1.46	0.43	0.85	0.75	0.87	0.75	1.75	0.17
Edu. (Postgraduate)	1.37	0.62	1.60	0.45	2.62	0.31	0.59	0.22	0.90	0.84	0.36	**0.036**	0.91	0.82
Freq. of Use (Daily)	1.47	0.64	2.02	0.31	0.08	0.09	1.97	0.27	3.33	0.15	2.16	0.15	1.85	0.22
Freq. of Use (Weekly)	1.13	0.84	2.27	0.12	0.35	0.22	0.78	0.50	0.83	0.65	1.70	0.16	1.04	0.90
Multi-user	0.63	0.42	0.73	0.50	0.14	**0.015**	1.05	0.90	0.63	0.27	0.68	0.27	0.79	0.47
Work use	0.65	0.47	0.61	0.30	2.49	0.27	0.98	0.97	1.04	0.92	1.30	0.48	3.26	**0.001**
Personal use	1.37	0.58	0.92	0.85	10.07	**0.011**	1.90	0.08	2.68	**0.019**	1.51	0.24	0.89	0.72
Health use	1.09	0.89	1.77	0.27	0.66	0.59	1.13	0.75	2.30	0.07	1.62	0.16	1.38	0.33

E.3 Perceptions of Plausibility

Table 8: Results for significant binary logistic regression models predicting whether participants found a given scenario to be plausible. As indicated by column names, the scenarios with significant demographic effects were the health advertising and database leak scenarios

	Health Advertising		**Database Leak**	
	$LL = -43.9$		$LL = -25.8$	
	OR	*p*	*OR*	*p*
Woman	0.54	0.30	1.68	0.51
Age (18-24)	0.15	0.18	0.00	1.00
Age (35 - 44)	0.12	0.08	0.00	1.00
Age (45 - 54)	0.02	**0.004**	0.00	1.00
Age (55+)	0.05	**0.035**	0.00	1.00
Edu. (HS)	0.30	0.22	0.32	0.35
Edu. (A-Level)	0.30	0.12	0.69	0.67
Edu. (Postgraduate)	1.84	0.52	1.94	1.00
Freq. of Use (Daily)	0.04	**0.002**	0.34	0.39
Freq. of Use (Weekly)	0.61	0.52	0.66	0.64
Multi-user	2.23	0.23	2.48	0.31
Work use	0.95	0.95	1.22	0.83
Personal use	0.27	**0.047**	1.01	0.99
Health use	0.82	0.75	11.50	**0.036**

E.4 Prior Awareness

Table 9: Results for significant binary logistic regression models predicting whether participants were aware of a given scenario prior to completing the survey. As indicated by column names, the scenarios with significant demographic effects were the model training, memorisation, law enforcement access, and deception scenarios.

	Model Training		**Memorisation**		**Law Enforcement Access**		**Deception**	
	$LL = -101.7$		$LL = -132.3$		$LL = -132.5$		$LL = -99.9$	
	OR	*p*	*OR*	*p*	*OR*	*p*	*OR*	*p*
Woman	0.76	0.45	0.65	0.18	0.69	0.24	0.66	0.27
Age (18-24)	0.48	0.20	0.42	0.10	0.78	0.62	1.03	0.96
Age (35 - 44)	0.88	0.79	0.50	0.08	0.45	**0.039**	0.70	0.47
Age (45 - 54)	0.38	0.07	0.45	0.09	0.27	**0.006**	1.04	0.95
Age (55+)	2.11	0.29	0.30	**0.027**	0.26	**0.016**	1.98	0.25
Edu. (HS)	0.21	**0.01**	0.41	0.14	0.89	0.83	2.89	0.07
Edu. (A-Level)	0.65	0.34	1.55	0.27	1.22	0.63	0.71	0.53
Edu. (Postgraduate)	1.86	0.26	1.67	0.21	0.78	0.52	1.32	0.56
Freq. of Use (Daily)	0.82	0.74	2.29	0.10	1.12	0.81	1.51	0.47
Freq. of Use (Weekly)	1.41	0.39	1.79	0.09	1.26	0.49	0.85	0.70
Multi-user	2.39	**0.029**	1.41	0.29	1.76	0.08	0.51	0.08
Work use	0.92	0.83	0.88	0.71	1.36	0.36	1.82	0.16
Personal use	0.97	0.94	0.78	0.43	1.67	0.11	1.48	0.32
Health use	1.08	0.84	0.78	0.46	0.75	0.38	1.60	0.22

E.5 Privacy Concern

Table 10: Results for significant binary logistic regression models predicting high levels of concern about scenarios. The only scenario with significant demographic effects was the deception scenario.

	Deception	
	$LL = -131.55$	
	OR	*p*
Woman	0.97	0.92
Age (18-24)	6.19	**0.003**
Age (35 - 44)	1.44	0.34
Age (45 - 54)	0.97	0.95
Age (55+)	1.95	0.21
Edu. (HS)	4.25	**0.036**
Edu. (A-Level)	0.75	0.48
Edu. (Postgraduate)	0.77	0.50
Freq. of Use (Daily)	0.63	0.35
Freq. of Use (Weekly)	0.52	0.06
Multi-user	1.00	1.00
Work use	1.17	0.64
Personal use	1.10	0.76
Health use	1.34	0.38