Understanding Users' Security and Privacy Concerns and Attitudes Towards Conversational AI Platforms

Mutahar Ali
University of California, Irvine
mutahara@uci.edu

Arjun Arunasalam
Purdue University
aarunasa@purdue.edu

Habiba Farrukh
University of California, Irvine,
habibaf@uci.edu

***Abstract*—The widespread adoption of conversational AI platforms has introduced new security and privacy risks. While these risks and their mitigation strategies have been extensively researched from a technical perspective, users' perceptions of these platforms' security and privacy remain largely unexplored. In this paper, we conduct a large-scale analysis of over 2.5M user posts from the r/ChatGPT Reddit community to understand users' security and privacy concerns and attitudes toward conversational AI platforms. Our qualitative analysis reveals that users are concerned about each stage of the data lifecycle (i.e., collection, usage, and retention). They seek mitigations for security vulnerabilities, compliance with privacy regulations, and greater transparency and control in data handling. We also find that users exhibit varied behaviors and preferences when interacting with these platforms. Some users proactively safeguard their data and adjust privacy settings, while others prioritize convenience over privacy risks, dismissing privacy concerns in favor of benefits, or feel resigned to inevitable data sharing. Through qualitative content and regression analysis, we discover that users' concerns evolve over time with the evolving AI landscape and are influenced by technological developments and major events. Based on our findings, we provide recommendations for users, platforms, enterprises, and policymakers to enhance transparency, improve data controls, and increase user trust and adoption.**

1. Introduction

Large language models (LLMs) [1], [2] have revolutionized artificial intelligence (AI), driving the widespread adoption of conversational AI platforms (e.g., ChatGPT [3], Gemini [4], and Claude [5]). These platforms allow users to interact with LLM-based AI systems through different modalities, such as text and voice. They are being integrated into a diverse range of applications, including customer service [6], [7], [8], personal assistants [9], [10], healthcare [11], [12], and finance [13], [14], due to their advanced natural language abilities, enabling more intuitive user interactions.

However, as conversational AI platforms become more popular, concerns surrounding their security and privacy (S&P) have also intensified. These platforms present unique S&P risks as their human-like conversational abilities can inadvertently encourage users to share sensitive information more freely than with traditional interfaces [15]. Prior work has also demonstrated that LLMs can memorize and reproduce sensitive information from their training data in response to malicious prompts [16], [17].

High-profile data leaks [18], corporate restrictions on platform usage due to suspected sensitive data exposures [19], [20], [21], and the drive for regulations (e.g., the White House Executive Order on AI adoption [22]) further highlight the critical nature of these risks.

While prior research has predominantly focused on the technical aspects of securing LLMs [23], [24], [25], [26], and in turn these platforms, there is a notable gap in understanding how users perceive and respond to these S&P risks. Preliminary studies have leveraged mixed methods (e.g., surveys, semi-structured interviews) to investigate users' S&P concerns about conversational AI platforms [27], [28]. These studies provide valuable initial insights into privacy harms and risks associated with AI systems, including common types of personal information users disclose and factors that influence trust in these platforms. However, these studies are conducted in controlled contexts with limited participant diversity and capture user concerns at a single point in time. Consequently, a comprehensive understanding of users' S&P concerns, behaviors, and preferences in interactions with conversational AI platforms across diverse user bases and how they change over time is needed.

To address this gap, we complement prior works by conducting a large-scale analysis of real-world, organic discussions on Reddit [29] to investigate the following research questions:

- **RQ1** What are users' S&P concerns related to conversational AI platforms?
- **RQ2** What are users' S&P attitudes toward conversational AI platforms?
- **RQ3** How do users' S&P concerns evolve over time, and how do major events in the AI ecosystem influence users' S&P concerns and attitudes?

To answer these questions, we collect and analyze a dataset comprising $\sim$ 2.5M posts ($\sim$ 180K submissions and $\sim$ 2.35M comments) from `r/ChatGPT`, the largest subreddit dedicated to conversational AI platforms, with over 7.4M members. Using keyword filtering and stratified sampling, we select and manually annotate a subset of $1,200$ posts from the entire corpus to fine-tune a RoBERTa-based

binary classifier [30] for identifying S&P-related content. Applying this classifier to the entire dataset yields 30,240 S&P-related posts. We iteratively sample and code a subset of these posts through qualitative analysis to develop a codebook. We then leverage an LLM-based multi-class classifier for post-hoc data labeling, assigning thematic codes to the entire S&P dataset. Finally, we conduct an interrupted time series regression analysis [31] to investigate how users' S&P concerns evolve over time, specifically in response to major events in the AI ecosystem.

Our qualitative and quantitative analysis reveals that users are primarily concerned about what personal and proprietary data these platforms collect and why (43.7%). They also worry about how the collected data is used (22.5%), particularly for training models or third-party sharing, and its retention (9.5%). Users also seek assurance in platform security and security of apps integrating conversational AI platforms (28.9%), compliance with privacy laws (9.6%), and desire transparent data handling practices and clear control over their data (11.1%).

We observe that users exhibit varied S&P behaviors and preferences when interacting with conversational AI platforms. Some are proactive in safeguarding their data, adjusting privacy settings, and limiting the information they share. Others are inquisitive, seeking information about how their data is collected, used, and stored. Some prioritize convenience and the benefits these platforms offer, often downplaying or dismissing privacy risks. Meanwhile, others feel resigned to data sharing as an unavoidable aspect of modern digital interactions.

Our longitudinal analysis further reveals that users' S&P concerns evolve over time, influenced by key events such as platform updates, regulatory changes, feature releases, and security incidents. For example, Microsoft's investment in OpenAI [32] sparked increased concern about third-party sharing, while Italy's temporary ban on ChatGPT [33] due to GDPR violations intensified discussions about compliance with data protection regulations.

Our study extends efforts in understanding users' S&P concerns and attitudes towards conversational AI platforms. It also highlights how these concerns shift in response to significant industry and regulatory events in the AI ecosystem. We conclude by synthesizing recommendations for key stakeholders, including recommendations for users to take proactive steps in managing their data privacy, platforms to improve transparency and create more intuitive data controls, enterprises to define clear usage guidelines, and policymakers to establish clear regulations and promote the standardization of privacy information. These recommendations collectively contribute to a safer and more trustworthy AI ecosystem.

In this paper, we make the following contributions:

- We conduct a large-scale analysis of more than 2.5M user posts from Reddit to understand users' S&P concerns and attitudes towards conversational AI platforms.
- We investigate how users' concerns evolve over time and identify the impact of major industry and regulatory events on users' concerns through regression analysis.
- We provide actionable recommendations for key stakeholders to support the safe, transparent, and responsible deployment and use of conversational AI platforms.

2. Related Work

Users' S&P Concerns and Attitudes. Prior research has explored users' security and privacy (S&P) concerns [34], behaviors [35], preferences [36], [37], and attitudes [38]. Studies have also explored users' S&P attitudes and concerns toward specific platforms and technologies. For example, studies on smart homes and IoT devices have highlighted user apprehensions about data collection and device security [39], [40], [41], [42], [43]. Similarly, studies on web browsers and online social networks have investigated users' views and satisfaction with privacy settings, uncovering a range of concerns about data sharing and control [44]. While prior research has examined general S&P attitudes or concerns toward specific technologies, we focus on the unique challenges and concerns that arise from the human-like interactions facilitated by conversational AI.

Online Discussions on S&P. Online discussion platforms (e.g., Reddit, X, StackOverflow) offer spaces for open dialogue, mutual learning, and peer support, providing rich, real-world data that reflects user concerns and behaviors. Researchers have leveraged online discussions to study S&P from different perspectives [45], [46], [47], [48], [39], [49], [50], [51]. For example, Li et al. [39] analyzed Reddit posts to investigate users' S&P considerations when adopting smart home technologies. Similarly, Tahaei et al. [47] conducted a qualitative analysis of privacy-related discussions on Stack Overflow to explore privacy challenges faced by developers. Wei et al. [46] studied user perceptions of targeted advertising by analyzing Twitter data. Several works have also investigated user privacy feedback by analyzing app reviews on app stores (e.g., Google Play Store) [52], [53], [54], [55], [56]. These studies illustrate the potential of online discourse to surface concerns and behaviors not easily captured through traditional surveys or interviews. Our work builds on this approach by analyzing discussions on the `r/ChatGPT` subreddit to capture nuanced and evolving user S&P perspectives unique to conversational AI platforms.

S&P in Conversational AI Platforms. Large Language Models (LLMs) [1], [2] are increasingly being leveraged to power AI systems that engage users in natural language across multiple modalities (e.g., text and voice). We refer to these AI systems as conversational AI platforms. These platforms vary in complexity, ranging from basic chatbots to intelligent agents with task completion capabilities, and have unique S&P challenges. LLMs' human-like conversational abilities can encourage users to share sensitive information more freely than with traditional interfaces [15]. Moreover, prior works show that LLMs can memorize and reproduce sensitive information from their training data [16], raising concerns about data leakage and privacy breaches.

A large body of work has explored the S&P of conversational AI platforms from a technical perspective [16],

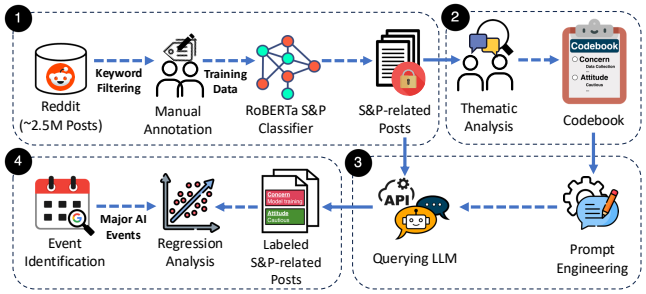


Figure 1: Methodology overview.

[17], [24], [25], [26], [57]. These studies primarily aim to understand vulnerabilities inherent to LLM architectures and behaviors. However, far less attention has been paid to the human factors associated with these risks.

Recent work by Lee et al. [58] studied AI practitioners' awareness, motivation, and ability to mitigate AI-related privacy risks, revealing limited awareness of AI-specific threats and insufficient incentives for privacy work. A few studies have also explored users' S&P concerns towards conversational AI platforms. Gumusel et al. [28] conducted semi-structured interviews with 13 participants to investigate privacy harms and risks in conversational AI. Zhang et al. [27] used a mixed-methods approach, analyzing a dataset [59] of user interactions with ChatGPT to categorize data disclosure scenarios in conversational AI and conducting interviews with 19 participants to investigate privacy concerns and disclosure behaviors. They identified patterns in user behavior and factors influencing trust during interactions.

While these studies provide valuable initial insights, they have three main limitations. First, they primarily involve participants from the United States, which may not represent the global user base of conversational AI platforms. Privacy concerns and regulatory landscapes vary across jurisdictions (e.g., GDPR [60] in Europe), leading to differing user concerns and behaviors. Second, these works focus on end-user interactions, overlooking S&P concerns from developers and enterprises deploying these platforms. Lastly, they explore users' perspectives in a controlled context at a single point in time, failing to capture how user concerns and attitudes evolve in response to new features and emerging risks.

In contrast, our work analyzes a large-scale dataset from Reddit, capturing a broader range of user concerns and attitudes over time. By focusing on discussions from a diverse user base, we capture real-world user interactions and include perspectives influenced by regional privacy regulations like the GDPR. Our analysis spans a significant time frame, allowing us to observe how users' S&P concerns evolve in response to major events such as new features, policy changes, and regulatory actions. This dynamic perspective offers a more comprehensive understanding of user attitudes toward S&P in conversational AI platforms.

3. Methodology

To understand users' security and privacy (S&P) concerns and attitudes toward conversational AI platforms, we conduct a multi-stage analysis of discussions on Reddit [29]. Reddit is an online discussion forum organized into topic-specific communities called subreddits, where users share, discuss, and vote on content in threaded structures. Each thread consists of a submission (the original post) and a series of comments. Reddit discussions represent large-scale, organic user-generated data that offer valuable insights into users' concerns, perceptions, behaviors, and information-seeking practices in the real-world [39], [49], [50].

Figure 1 presents an overview of our methodology, which involves four main steps. We begin by collecting a large dataset of Reddit posts and annotating a subsample to train a classifier that identifies content related to S&P ❶. Next, we conduct a qualitative analysis to identify users' S&P concerns and attitudes towards conversational AI platforms ❷. Building on this, we leverage an LLM-based multi-class classifier to label all S&P-related posts with thematic codes derived from our qualitative analysis ❸. Finally, we perform a regression analysis to investigate how users' S&P concerns shift in response to major events in the AI ecosystem ❹.

3.1. Identifying S&P-related posts

To understand users' concerns and attitudes, we first compiled a list of popular subreddits related to conversational AI platforms using Reddit's search engine [29], as shown in Appendix A Table 1. From this list, we selected `r/ChatGPT` for our analysis based on several reasons. First, it is the largest and most active subreddit focused on conversational AI with 7.4M members - more than 4 times the size of the next largest subreddit (`r/OpenAI`) - and ranks $86th$ among all Reddit communities. Second, it has high user engagement and activity ($\sim$ 300 daily posts and $\sim$ 4,000 comments). Third, despite its branding, `r/ChatGPT` regularly features discussions that extend beyond ChatGPT. For example, over 7,500 posts mention local LLMs (e.g., LLaMA [61]), and more than 43,000 reference other commercial platforms like Gemini [4] and Claude [5]), indicating a broader topical scope. Lastly, our qualitative analysis shows that the subreddit hosts diverse user attitudes (Section 5), ranging from cautious and curious to privacy-dismissive and resigned. Therefore, selecting `r/ChatGPT` allows us to capture substantial data covering diverse S&P concerns across different conversational AI platforms.

To conduct our analysis, we gathered all posts from `r/ChatGPT` from its inception in December 2022 till July 2024, using publicly available Reddit dumps [62]. After removing deleted posts, bot-generated content [1], and duplicate entries, our final dataset included $\sim$ 2.5M posts ($\sim$ 180K submissions and $\sim$ 2.35M comments, collectively termed "posts").

Given the dataset size, identifying S&P-related discussion threads is a significant challenge. An intuitive approach is to perform a search with keywords associated with S&P

1. Bot-generated posts are identified using known bot signatures, such as the presence of the word "bot" in the username or phrases like "Beep boop, I'm a bot" [63].

to surface relevant posts [49]. However, this approach can suffer from low coverage, missing relevant posts that use alternative phrasing, and high false positives, especially in communities like `r/ChatGPT` where users frequently share generative content (e.g., fictional stories, essays, or poems) that may include keywords in unrelated contexts. To address these limitations, we adopt a hybrid approach combining keyword filtering with machine learning.

Keyword Filtering and Sampling. We began by reviewing the literature on S&P issues in conversational AI and AI risk taxonomies [16], [27], [28], [64], [65], [66], [67], [68], [69] to compile an initial list of candidate keywords. We refined this list iteratively by searching Reddit posts, reviewing matched results, identifying new keywords, and updating the list accordingly. Using this iterative approach, we compiled a list of 118 S&P-related keywords and expressions, presented in Appendix B Table 3.

To ensure balanced representation and reduce bias toward frequently mentioned keywords, we grouped our keywords into 6 thematic categories based on prior works [66], [70], [71]. We then performed stratified random sampling, selecting 100 posts from each group, resulting in 600 posts potentially related to S&P. To enable the classifier to distinguish S&P-related content from other topics and to capture themes outside our keyword list, we complemented this sample with 600 randomly selected posts that did not contain any keywords. In total, our seed corpus consisted of $1,200$ posts, balanced between submissions and comments.

Manual Annotation. To construct a reliable seed corpus for training our classifier, we manually annotated the 1,200 sampled posts, labeling each as either S&P-related or not S&P-related. To ensure consistency and minimize subjectivity, two authors (experts in information S&P) jointly reviewed an initial subset of 200 posts to create a detailed annotation guide, defining clear boundaries between S&P and non-S&P content. Using this guide, the two authors then independently labeled the corpus, achieving high inter-rater agreement (Cohen's Kappa, $\kappa = 0.82$). The final labeled dataset included 209 S&P-related posts (17.4%) and 991 non-S&P posts.

S&P-Related Post Classification. To identify S&P-related posts within our dataset, we fine-tuned a binary classifier using RoBERTa [30], a pre-trained language model for text classification tasks. Given the nuanced nature of S&P topics and the requirement for both high precision and recall, we selected RoBERTa based on its superior performance over other models (i.e., DeBERTa [72], GPT-4 [73] and Gemini Flash 1.5 [74]), as shown in Appendix C Table 2. The classifier achieved an accuracy of 96% and an F1-score of 82%. Applying this classifier to the entire dataset ($\sim$ 2.5M posts), yielded 30,240 S&P-related posts (1.2%), authored by 18,851 unique Reddit users. Appendix D Figure 5 shows the weekly counts of S&P-related posts over time.

3.2. Data Analysis

3.2.1. Sampling and Coding. To identify users' S&P concerns and attitudes, we conducted a thematic analysis. We first developed an initial codebook informed by our keyword exploration and manual annotation for S&P classification (Section 3.1). To refine the codebook, two authors iteratively sampled and coded S&P-related posts over two weeks until reaching thematic saturation [75] at 440 posts.

These 440 posts were authored by 433 unique Reddit users. With each sample, we updated our codebook by introducing new codes for emerging concerns and attitudes. We then grouped similar codes into broader themes that align with specific stages of the data lifecycle (data collection, usage, and retention) as well as broader concerns that span multiple stages (security vulnerabilities, regulatory compliance, and transparency and control). We do not present intercoder agreement as all posts were reviewed together [76].

3.2.2. Prevalence Analysis. To analyze the prevalence of concerns and their evolution across events (Section 6), we labeled the entire dataset of S&P-related posts (30,240 posts) using an LLM-based multi-class classifier. Given the scale of our dataset, manually annotating all posts was infeasible. Therefore, similar to prior works [50], we leveraged LLMs for post-hoc data labeling to label each post with codes from our qualitative analysis (Section 4). Specifically, we used GPT-4o [73] with a hierarchical prompting [77] approach. First, the model predicted the high-level concerns (themes) e.g., data collection, data usage. Then, for each predicted theme, we used a separate prompt to predict the low-level concerns (sub-themes) e.g., personal data, model training. We evaluated the classifier on our dataset of 440 manually coded posts (Section 3.2.1), achieving an average accuracy of 96.8% and F1-score of 91.2%. Detailed performance metrics for each theme and sub-theme are presented in Appendix E Table 4. After applying the classifier to the entire dataset, we conducted a secondary evaluation to assess its performance on unseen data. We randomly sampled 100 LLM-classified posts and manually coded them to generate ground truth labels. The classifier achieved an accuracy of 98.5% and an F1-score of 95.1% in labeling these posts.

3.3. Ethical considerations

We strictly follow ethical guidelines for research involving publicly available online data. All data used in this study was collected solely for non-commercial academic purposes. Before conducting this research, we consulted with the Institutional Review Board (IRB) at our institution. The study was deemed to be IRB exempt, as it involved the use of publicly available information.

To protect user privacy, all usernames have been anonymized, and no personal identifiers have been retained in our dataset. The data is used exclusively for internal analysis and has not been redistributed. Additionally, we do not release any derivative products, such as our classifier for identifying S&P-related posts, to ensure the privacy of the individuals represented in the data.

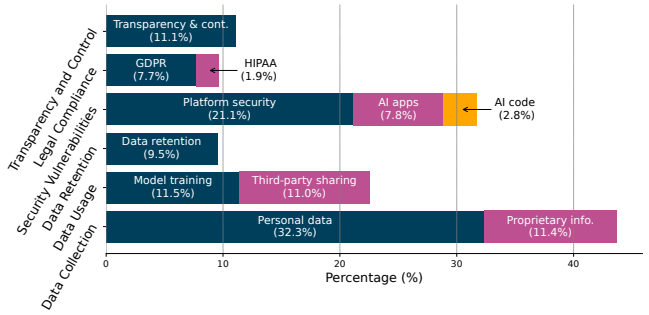


Figure 2: Prevalence of users' S&P concerns.

4. RQ1: Users' S&P Concerns

Our analysis reveals a range of recurring security and privacy (S&P) concerns regarding conversational AI platforms (e.g., ChatGPT). We categorize these concerns into six main themes: (1) data collection, (2) data usage, (3) data retention, (4) security vulnerabilities, (5) legal compliance, and (6) transparency and control. Figure 2 presents the prevalence of each theme and its sub-themes. These prevalence metrics can guide platforms in prioritizing improvements by highlighting the most common user concerns e.g., the need for greater transparency and control around data collection and model training. Below, we detail our findings for each theme.

4.1. Data Collection

Users are wary of the nature and scope of data collected by conversational AI platforms. We discover concerns related to *personal data* – information that identifies individuals or reveals private details, and *proprietary information* – sensitive corporate information and intellectual property, e.g., business plans, trade secrets, ideas, and work products.

4.1.1. Personal Data. Users discuss that interacting with conversational AI platforms often requires sharing personal data, either directly or via passive data collection.

Personally Identifiable Information (PII). PII refers to any data that can uniquely identify an individual [78], such as *"IP address, email, username, phone number, and device ID"*. Users express concern regarding PII collection by conversational AI platforms. For instance, one user expressed discomfort about the AI platform having access to their name and university information: *"The data gathering is scary, it know[s] my name and my university"*. Some users review the privacy policies of popular AI platforms to understand their policies regarding PII collection. For example, one user pointed out that ChatGPT's privacy policy *"[states] that they can personally identify you"*. Another user mentioned that *"they can put the data and the identity info together"*.

Activity and Behavioral Data. Users are concerned that conversational AI platforms may track and monitor various aspects of their digital behavior beyond conversational data. Their main concerns revolve around the collection of browsing activity, GPS location, and data from other apps on the user's devices. Users report instances where the AI response includes personal details that they did not provide, questioning whether the platform is monitoring their online activity and browsing habits. For example, one user noted, *"Does ChatGPT have access to our browser & site visits? It figured out my [university's] name despite [me] not previously mentioning it."*

Users also worry about AI platforms accessing their location data without permission. For instance, one user noted ChatGPT's apparent knowledge of their location: *"Does ChatGPT really have GPS location without enabling? I never told it my location before. It guessed down to the [right] town."* In contrast, some users do not find this surprising, comparing it to how Google Maps works (e.g., *"Do you freak out when you Google 'Food near me' and it actually searches restaurants in your area?"*).

4.1.2. Proprietary Information. Several discussions in our analysis involve employees in corporate settings expressing concern about using conversational AI platforms with confidential and proprietary information such as business plans, internal emails, trade secrets, project details, and software. Users also express concerns about the ownership and privacy of their intellectual property (IP).

Corporate Data. Many users view the use of AI tools in professional environments as a *"corporate nightmare"* due to the risk of exposure of sensitive data to unauthorized parties. Some users worry that this may lead to *"consequences"* such as legal repercussions or suspension e.g., *"[ChatGPT] does not treat submitted info as confidential ... If anyone finds out you are using [it] in your day job ... you are probably in material breach of your employment contract."*.

Many users are also wary of the privacy risks of using conversational AI with work-related information, especially in fields bound by confidentiality, such as legal, finance, and consulting roles. For example, one user expressed their concern regarding the use of AI tools in the legal profession: *"The information you ask from [the AI] or give [it] access to is most likely not private. Isn't this a [major issue] in the law space [with] client confidentiality [being] broken?"*.

Software. Many users share specific concerns about the privacy of code and software shared with platforms, expressing skepticism about the *"settings to disable data collection"* and questioning how they can *"prevent ChatGPT from spreading [their] information"* to keep sensitive code private. For instance, a user inquired, *"When using GPT, is there a way to keep the code/data you are feeding it private, like if I am working with patented/proprietary code and do not want it exposed outside my company?"*

To address S&P concerns, some conversational AI platforms offer specialized tiers for corporate use [79], [80] with added S&P features (e.g., exclusion of conversation data from model training). Nevertheless, users remain cautious and uncertain about the actual privacy protections these plans offer. For example, one user expressed concern about risks associated with ChatGPT Teams after reviewing OpenAI's privacy policy: *"The data in Team tier is not used for training, but can still be viewed by other people ... what's worse [is that] access is not limited to OpenAI staff"*.

Ideas and Creative Works. In addition to corporate concerns, users who turn to conversational AI platforms for creative tasks (e.g., brainstorming, refining startup concepts) worry about these platforms *"stealing"* or misappropriating their original ideas. For instance, one user questioned whether sharing a story draft with ChatGPT might result in the story being misused, asking, *"If I had a story rough draft and I tell it to ChatGPT to get it refined, will the story end up getting stolen?"* Another user voiced suspicion that ChatGPT might be registering trademarks on their ideas, advising caution against using AI tools for creative tasks: *"I'm very suspicious that ChatGPT is stealing my ideas [from] brainstorming and getting trademarks or copyrights on certain names or ideas before I get a chance to. Beware of using these tools for any personal projects or ideas or inventions."*

4.2. Data Usage

Concerns about data usage revolve around its use for improving AI models and sharing with third parties.

4.2.1. Model Training. We find that users have two primary concerns regarding the use of their data in these training processes: the involvement of human reviewers in evaluating user data for training and the potential for models to "memorize" and disclose specific details from user interactions.

Human Review. Human oversight in model training frequently involves annotators reviewing user interactions to label data, guide improvements, and ensure accuracy [81]. Users express concerns that PII might be exposed during model training: *"When an AI system is in beta testing, the inputs and outputs of the system may be read by humans who are helping to train the model. This means that your PII could potentially be seen by these humans."*

We also find specific concerns related to popular conversational AI platforms. One Bard (now Gemini [4]) user advised against entering any private information in Bard, recalling Bard's explicit notice: *"Your conversations are processed by human reviewers to improve the technologies powering Bard."* Some users point out that ChatGPT lacks a similar warning, potentially leaving users unaware that human reviewers may access their interactions. For example, one user stated, *"At least Google "says" not to [share private information]. While [its] competitor loves binge eating data."*

Memorization. Prior work has shown that LLMs can *"memorize"* and reproduce specific pieces of information from their training data [16]. Users worry that sensitive information entered into conversational AI platforms could inadvertently be disclosed to other platform users through the models' responses. For example, one user cautioned that *"[ChatGPT] learns from the data"* while another noted that data can be *"regurgitated to other users in some form."* Users feel this is a major issue, as memorized data could include PII. For instance, one user described how ChatGPT *"doxxed"* a random person in its response: *"I sent ChatGPT a text about charisma I wanted him to summarize ... [it wrote] out [the] most personal and private information of a random man."*

4.2.2. Third-party Sharing. Users share concerns about who may access their data, for what purpose, and under what conditions. For instance, one user asked, *"Does OpenAI share user activity with third parties? And if so, which parties?"*

Affiliates. Many users are skeptical of data sharing within corporate ecosystems, where affiliated companies might gain access to their information across interconnected platforms. This concern is highlighted by users of conversational AI platforms developed by large tech conglomerates. For instance, one user questioned whether Meta AI was *"trained on Whatsapp chats"*, while another speculated that Microsoft CoPilot *"already has [the] data"* from Outlook and Teams. Some users expressed concern regarding data sharing between Microsoft and OpenAI: *"OpenAI [can] pump all that juicy data to Microsoft while tying it to your device."*

Data Brokers and Advertisers. Another major concern for users is the potential sale of their personal information to data brokers and its use for targeted advertising [82]. For example, one user expressed their discomfort after reading OpenAI's Terms of Service, stating that *"[they allow] *all* personal information to be sold to any third party for any reason they think is OK ... specifically including advertisers and marketing services."* Another user shared a similar worry: *"[ChatGPT's privacy policy] literally says they will sell [users'] personal data."*

Users foresee various privacy consequences of sharing data with brokers and advertisers. Some worry that targeted ads could manipulate their purchasing choices, while others express deeper concerns about personal data being shared with insurance companies or lenders, possibly affecting loan eligibility or insurance premiums. For instance, one user commented, *"I'm more concerned about when advertisers get to influence training data. Google search is heinous enough at this point; imagine when it's even more subtle."* Another user noted, *"I would be more worried that my information [could] be sold to insurance brokers or lenders and that [may impact] my ability to get certain insurance or credit. Who would want to lend someone money for a house or give a suicidal person an insurance policy?"*

Government Authorities. Conversational AI platforms may be legally obligated to disclose user data to government authorities in certain situations, such as when a conversation includes references to imminent harm or in response to a court order or subpoena [83]. This raises concerns among users about the privacy of their interactions with the platforms. For example, one user highlighted the ease of government access, noting, *"If the US government subpoenas data you sent to OpenAI (including prompts and their responses) then OpenAI will almost certainly and immediately comply."*

4.3. Data Retention

User concerns around data retention mainly center around incomplete deletion, where data appears hidden rather than fully erased, and data storage practices.

Incomplete Data Deletion. One common perception among users is that after deletion, data is removed from the user's

view but persists on the backend. For instance, one user stated that *"when you delete a conversation, ChatGPT actually hides it instead of deleting it."* Another user raised suspicion, stating *"I've said some stuff that I would rather not be seen by anyone. I've already click[ed] the delete button but is it really deleted?"* Commenting on OpenAI's Terms of Service (ToS), another user pointed out that *"nothing in their ToS says they will remove content ... it's marked as deleted, and then hidden from the user."* These posts reflect a broader sentiment of mistrust regarding data handling practices among conversational AI platforms. One user summarized this sentiment: *"There is no way to know they will actually delete your data. If OpenAI was nonprofit, their claim would be more believable."*

Data Storage Practices. Another user concern is that personal data is implicitly stored within model weights after training, making it nearly impossible to erase. One user noted: *"OpenAI has used your personal data to train their model and cannot easily remove it."* Some users are also concerned that this may violate data protection laws (e.g., GDPR [60]) under which *"companies must let you delete your personal data"* (See Section 4.5.1). Moreover, users question whether platforms retain data for extended periods in ways that are inaccessible or unverifiable (e.g., *"We don't know if OpenAI archives the data or actually deletes it but my guess is they archive as much of it as they can."*)

4.4. Security Vulnerabilities

Apart from concerns regarding data collection, usage, and retention, users are also concerned about the security of conversational AI platforms and the data shared with them.

4.4.1. Platform Security. We find that users frequently express concerns about the security of conversational AI platforms, especially the risk that sensitive data might be exposed due to software bugs and data breaches.

Software Bugs. Users worry that software bugs in AI platforms may have security implications, potentially exposing sensitive user data to unauthorized parties. A notable example occurred in early 2023 when a bug in ChatGPT allowed users to view the titles of other users' conversations [18], which one user described as *"a massive privacy problem."* Another user, alarmed by seeing unknown conversation titles in their history, decided to delete their account, fearing that their personal information was compromised: *"I believe my information is shared by someone or mixed up with [another] person's information ... As soon as I am being refunded, I plan to delete my account"*. Some users believed that their account was *"hacked"* after seeing unknown conversations in their chat history.

Users also specifically discuss the security of ChatGPT's code interpreter, fearing how vulnerabilities could be exploited to steal data. One user suggested that it can be used to steal user data by getting the user to *"[paste] a malicious URL into ChatGPT"* and using it to *"run instructions"* in the browser to *"grab data and send it to a third-party server."* Such concerns emphasize users' apprehension about the security robustness of these platforms.

Data Breaches. Another significant user concern involves the risk of data breaches, which could expose private data from stored conversations. For instance, a user noted, *"I'm scared that there's going to be a data breach and my information will be leaked."* This apprehension makes users cautious about sharing personal details with AI platforms, as highlighted by a user who warned: *"Be careful with the info you provide ChatGPT ... If the system is hacked or experiences a data breach, the PII you provide could be accessed and potentially misused by unauthorized parties."* Another user explained how, even with chat history disabled, data is *"retained for a time"*, leaving it susceptible to potential data breaches.

4.4.2. AI-enabled Applications. Conversational AI is increasingly being integrated into applications such as e-commerce and travel booking platforms. However, users who leverage these systems are concerned about the security of their products and services.

Prompt Injections and Jailbreaks. Users worry about the susceptibility of their chat services to jailbreaks [84] and prompt injection attacks [85]. They fear that these attacks could enable unauthorized access to sensitive data or allow users to bypass normal procedures in consumer-facing applications. For instance, one user wrote: *"A lot of the companies with early integration are susceptible to prompt injections."* Another user queried: *"Say I wanted to make a chatbot for a business using the API, would I be able to prevent a user from feeding it prompts that would significantly change the output (like DAN for example)?"*

Custom GPTs. The recent introduction of customizable GPTs [86], which users can configure with custom instructions, data, and API access and make available to other users via the GPT store [87], has also fueled concerns among users. Users are concerned that adversaries could manipulate these custom GPTs to expose private data (e.g., custom instructions). For example, one user noted: *"Even after telling my custom GPT to not leak the information from the knowledge database and the custom instructions through which it is trained, it's still revealing those things."* Another user described how custom GPTs can be tricked into *"spilling secrets"*, including *"sensitive documents"*.

4.4.3. Security of AI-generated code. Although several users acknowledge the utility provided by many conversational AI platforms (e.g., ChatGPT) that enable code generation, debugging, and even direct code execution through built-in interpreters, we find that many developers and professionals express concerns about security vulnerabilities present in the generated code, describing it as *"bug filled"*. They worry that the code generated by these platforms does not *"follow up to date security principles"*, and many recommend against its use in software systems that handle sensitive data. For example, one user shared: *"[I am scared] at the idea of some startup using [AI-generated code] excessively for something that stores and handles sensitive data."* Similarly,

a security professional expressed his concern about the scale at which AI-generated code is deployed, stating that *"Society is absolutely unprepared for the knock-on effects of releasing all this [insecure] code into production ... there are just not enough of us [security researchers] to handle this."*

Human vs. AI-Generated Code. Users argue that over-reliance on AI-generated code inadvertently leads to vulnerabilities in software systems. For example, one user commented: *"This is how you get zero-day vulnerabilities in your microservice. Anyone relying on completely machine-generated code to save a few thousand dollars is a [fool]."* Interestingly, some users counter-argue that human code may contain similar vulnerabilities (e.g., *"Hiring human programmers is also an excellent way to get vulnerabilities in your code, though."*)

Persistent Vulnerabilities. Some users note that AI-generated code can perpetuate security bugs if they are common in the training data. For instance, one user described that even after a vulnerability in AI-generated code is spotted, the platform may continue to generate code with the same vulnerabilities due to its presence in the model's training data: *"What happens when a hacker finds a vulnerability in AI-generated code but the [LLM] keeps re-creating the same vulnerabilities because the code is so common"*.

4.5. Privacy Regulations Compliance

Our analysis shows that many users demonstrate awareness regarding data and privacy protection laws, particularly the General Data Protection Regulation (GDPR) [60] in the European Union (EU) and the Health Insurance Portability and Accountability Act (HIPAA) [88] in the United States.

4.5.1. GDPR Compliance. GDPR [60] is an EU law that protects individuals' personal data, setting strict guidelines for data collection, storage, and user consent. Users are concerned that conversational AI platforms may not fully comply with GDPR requirements. For instance, one user flagged the presence of personally identifiable information (PII) in the AI platform's training data, stating, *"Training data has PII, which breaks GDPR."*

Under GDPR, individuals have a *"right to be forgotten"*, which requires organizations to delete all user-associated data upon request [60]. Users express skepticism on conversational AI platforms' ability to meet this requirement, particularly given that user data might be embedded in the AI model itself (as discussed in Section 4.3). For instance, a user noted, *"ChatGPT doesn't really forget data the same way as erasing under the EU right to be forgotten requires ... once a chat is used in machine learning, you can't ever truly erase it"*.

Users also discuss GDPR's restrictions on collecting data from minors and question whether AI platforms implement proper age controls to comply with these guidelines. For example, one user stated that there is *"no age limit for minors"*. Another user wrote that the technology cannot *"target"* children if it is to be compliant with GDPR.

4.5.2. HIPAA Compliance. HIPAA is a U.S. law that mandates the protection and confidential handling of sensitive patient health information by healthcare providers and associated entities [88]. Remarkably, many users are not only aware of HIPAA's strict data protection requirements but also actively discuss how conversational AI platforms may fall short of these standards. Despite HIPAA's protections, users report instances they perceive as non-compliant. One user, for instance, shared that they have observed *"multiple doctors"* using ChatGPT with sensitive medical data, raising alarm over potential HIPAA violations. Another user expressed concern over healthcare providers potentially uploading patient records to ChatGPT, remarking, *"I really hope that you are not uploading others' medical records to ChatGPT. That would be a massive PII and HIPAA violation."*

Users are similarly apprehensive of using conversational AI platforms in therapeutic contexts. They worry that, unlike interactions with licensed therapists, AI interactions are not governed by confidentiality rules (e.g., *"If you talk to a psychologist, there are rules ... [However,] everything you write to ChatGPT is free to use for retraining"*).

4.6. Transparency and Control

Users frequently report that the transparency and control measures in conversational AI platforms are inadequate or misleading, highlighting that the options provided are unclear, limited, or even manipulative.

Data Controls. Many conversational AI platforms offer users some control over data sharing, including options to manage data usage for model training [89]. However, users report significant limitations and confusion surrounding these controls, which undermines their ability to manage their data.

Many users are concerned that data sharing for model training is typically enabled by default, requiring users to manually opt out. For instance, one user noted, *"There's an option under settings on GPT-4 that's automatically checked yes for them to use your data for model training purposes."* Users also discuss how data control options are unclear and challenging to navigate, which creates confusion about how to effectively disable data sharing for model training purposes. For example, one user stated that *"users cannot easily opt out of data sharing"* in ChatGPT. Another user reported how the setting to opt out of sharing in ChatGPT was renamed, but none of the guides were updated accordingly: *"It used to say 'chat history and model training.' Now, it only says to 'improve the model for everyone'."*

Dark Patterns. Many users believe that AI platforms employ dark patterns to subtly nudge them toward sharing data by limiting functionality or providing minimal privacy options. Users report functionality issues or software bugs when they opt out of data collection, which they feel pressures them into opting in. For instance, one user questioned if UI bugs (e.g., broken scroll) were *"intentional on OpenAI's part to cripple your experience if you don't want to share data for training?"*. Similarly, users wonder if opting out of data sharing prevented them from *"getting access to the voice [chat] feature"* or *"logging into [the ChatGPT] iOS app"*.

Users subscribing to paid plans for conversational AI platforms feel especially cheated and penalized for prioritizing privacy. For instance, one user noted, *"I just paid for [ChatGPT Plus] and [I] can't believe that I have to keep chat history enabled to use [the] code interpreter, meaning I can't use any sensitive data for data analysis."*

Policy Changes and Consent. Users also express concerns about AI platforms' frequent and often confusing policy changes, especially when these updates affect default privacy settings or alter permissions without clear communication. Many users highlight inconsistencies in privacy information across official sources. For instance, one user stated *"[ChatGPT's] website contains references to both the new and old [data] processes[,] making it hard to understand"*. Another user mentioned that the website lacks transparency and *"clear notices"*, especially regarding whether users can fully opt out of data collection.

Some users discuss that policy changes may be designed to exploit default settings. For example, one user reported that a recent ChatGPT policy update switched their data-sharing preference to *"yes"* even though they had previously opted out. Many users express frustration with being coerced into accepting new terms, as declining updates often prevent them from using their accounts altogether. For instance, one user noted, *"If you do not agree to the updates, you may delete your account. But this sounds a bit too much like 'if you don't like this, you can go and f*** yourself'."*

Key Takeaways (RQ1). Our analysis highlights users' strong concerns around data privacy, security, and control in conversational AI platforms. Users are uneasy about extensive data collection, unclear data usage, and the permanence of stored information, particularly fearing inadequate protection of personal data. Despite transparency and control features, many find these insufficient or confusing, which intensifies mistrust. These concerns are further fueled by potential noncompliance with privacy and data regulations.

5. RQ2: Users' S&P Attitudes

Our analysis reveals that users can be categorized into four groups based on their S&P attitudes (i.e., behaviors and preferences) towards conversational AI platforms: *cautious* – aware of privacy risks and taking proactive steps to protect their data, *inquisitive* – actively seeking information, *privacy-dismissive* – indifferent to privacy risks, and *resigned* – passively accepting privacy risks.

5.1. Cautious Users

Cautious users are highly aware of the privacy risks associated with conversational AI platforms. They prioritize data protection and actively implement measures to safeguard their information, often sharing tips and strategies with others. Their behaviors include adjusting privacy settings when using online AI platforms, opting for local solutions, and incorporating guardrails in AI-enabled applications.

Securing Interactions with Online Platforms. To mitigate privacy risks from cloud-based conversational AI platforms, cautious users take several actions. They restrict sensitive data in their prompts by sanitizing inputs to remove personally identifiable information (PII). Some users manually scrub their input text before sending it to the AI, as one suggested: *"You should scrub input text before sending it to the LLM."* Others seek tools to automate this process or use hypothetical or anonymized data. For example, a user mentioned: *"I also sanitize and anonymize private information by running the data through a Python script."*

Many cautious users use data control options to disable training on their conversations. For instance, one user noted: *"Turning off chat history and model training can be done to protect your privacy and ensure confidentiality."* Others regularly delete chat histories or even their accounts to prevent data retention and distribution (e.g., "I do not want my private information shared with someone online.").

Using Local Solutions. Believing local solutions to be more secure and private than cloud-based ones, many cautious users prefer open-source LLMs (e.g., Llama [90]) that operate on their devices. For instance, one user stated: *"An on-device AI is secure, more private, and provides more scope for personalization."* Another asked for offline models to avoid data interception risks: *"I worry that others can intercept my interaction ... is there a private, offline version?"* Users also report using local models in practice, removing reliance on third-party cloud infrastructure and maintaining direct control over their data. For instance, a user recommended: *"If you want to use such generative models with sensitive data, I recommend that you use [an] open-source model like LLaMA or Mistral and deploy [it] on-premise."*

Implementing Guardrails. Cautious users extend their privacy-preserving practices to designing their own AI services. They stay informed about new attacks and implement guardrails to prevent exploitation. For example, one user noted: *"I am creating a list of malicious prompts and different jailbreaking methods so it can identify when someone is trying to gain access."* Another user designed a prompt to stop custom GPTs from sharing private instructions: *"I've come up with a prompt which you can add to the end of your custom GPT instructions to protect it."*

5.2. Inquisitive Users

Inquisitive users actively seek to understand the privacy and data management practices of conversational AI platforms. They frequently question specific functionalities, such as chat history and training settings, to assess how their data is used. These users often balance concerns about privacy with the desire for convenience and efficiency in their tasks.

Engaging with the Community. Inquisitive users often turn to others in the community to gather information and perspectives on S&P practices. They ask questions to make informed decisions, such as: *"As OpenAI uses the data entered into GPT models to retrain them, would you be worried that the data you put in could be compromised if this data becomes exposed? Is this an issue that bothers you personally/with your work?"*, *"If I'm paying for Copilot*

Pro, will Microsoft use [my data] to personalize ads and sell/transfer [it] to third parties?", or *"Do you have chat history and training enabled?"*

Consulting AI Agents. Some users directly ask AI agents about their security practices to gain insights. Some users trust the AI's responses. For instance, they asked the platform: *"Is [ChatGPT's] chat history really deleted?"* and received an affirmative answer. Others remain skeptical, stating: *"ChatGPT stores information about its users [but] it will prefer to reject this fact if you ask."*

5.3. Privacy-Dismissive Users

These users exhibit indifference to privacy risks, prioritizing the benefits of AI platforms. They often criticize privacy regulations, which they perceive as barriers to technological advancement, and may dismiss others' S&P concerns.

Prioritizing Benefits Over Risks. These users believe the advantages of conversational AI outweigh potential privacy risks, adopting a pragmatic approach that values immediate utility. They highlight how existing devices are already conducting surveillance, e.g., *"We are exchanging modern convenience of technologies that make our lives easier for privacy. Every single device you buy or website you visit is already doing surveillance... why would I actually care?"*

Criticizing Privacy Regulations. Privacy-dismissive users often advocate for more lenient privacy laws, viewing regulations like GDPR as overly restrictive and hindering innovation. For instance, one user noted: *"The obsession with privacy in the EU is both hilarious ... Who cares if some random person overseas knows some of my personal data?"* They feel disappointed when privacy laws delay access to new features e.g., ChatGPT's memory feature [91], which was delayed *"because of privacy rights in the EU."* Another user voiced frustration about not being allowed to waive privacy rights for early access: *"As an EU citizen, I really hate that they don't even allow me to opt in."*

Dismissing Others' Concerns. Privacy-dismissive users may also overlook or trivialize others' concerns. For instance, one user responded to privacy worries by saying: *"If you don't like it, don't use it. Who cares if they have your data? ... I'd give them all my data to improve the bot if I could!"* Similarly, they dismiss concerns about targeted advertising: *"If it helps you change your thought patterns in a positive way, who cares what ads are tailored to you because of it?"*

5.4. Resigned Users

Resigned users feel overwhelmed by the pervasive nature of data collection and perceive resistance as futile. They believe that engaging with technology inevitably involves sacrificing privacy, expressing a sense of inevitability and helplessness regarding data privacy. For example, one user stated: *"there is no running [from it]. [It's] scary to believe that the AI has the potential to access all this info about us as we use it."* Another user noted: *"At some point, I gave up. If I want the latest tech, I have to sell my soul/privacy."*

These users argue that digital privacy is an illusion, as their data is collected regardless of their use of AI platforms. For example, one user noted: *"[Your] data is being harvested no matter what you do if you have a phone so who cares"*

Key Takeaways (RQ2). Our analysis reveals four S&P attitudes towards conversational AI platforms. Cautious users actively implement safeguards to protect their data, while inquisitive users seek to understand and navigate data usage policies. Privacy-dismissive users prioritize convenience, often criticizing strict privacy regulations, and resigned users feel that privacy is a lost cause in the digital age.

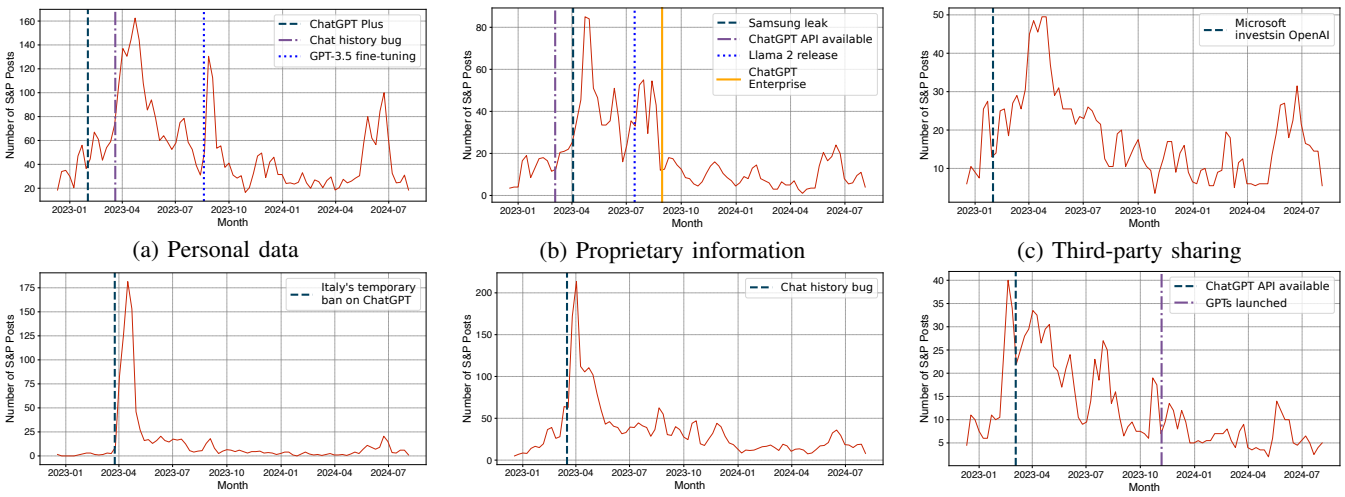
6. RQ3: Longitudinal Analysis

To understand how users' S&P concerns evolve over time and change in response to specific events, we first compiled a list of significant events related to conversational AI platforms. Two authors conducted a manual review of Google search results (from November 2022 to July 2024) using relevant keywords from our qualitative analysis (Sections 4 & 5). By cross-referencing official announcements, we identified 38 relevant events, including major announcements, platform and model launches, and feature releases. Our goal is not to compile an exhaustive list of events but to understand how different event types may impact different user concerns. Appendix F Figure 4 shows the timeline of compiled events.

We conducted an interrupted time series regression analysis [31] to identify events significantly impacting S&P discussions. We used an Ordinary Least Squares regression model to predict daily post counts within a seven-week window before and after each event. The model included time (in days), capturing the overall trend in the number of discussions, an intervention term (binary variable: 0 before and 1 after the event, measuring immediate effects), and an interaction term (time x intervention) for post-event trend differences. We used a p-value of 0.05 with Bonferroni correction [92] for multiple comparisons and manually validated the model's findings by inspecting discussions surrounding each event. Figure 3 illustrates how major events influence S&P discussions.

Launches and Acquisitions. We find that launch and acquisition events impact users' discussions about data collection and usage, specifically proprietary information and data sharing with third parties. For instance, when Microsoft invested 10B dollars in OpenAI [32], many users discussed the investment as an acquisition of OpenAI, considering Microsoft in control of OpenAI's data control policies. We also observe a peak in users' concerns about third-party data sharing (Figure 3c) after the event. Several users noted concerns about data monetization and targeted advertising. For instance, one user mentioned *"Microsoft is a company which is hellbent on making money after spending money ... [They may implement] ads-based revenue like Google."*

Another significant event that raised users' concerns about proprietary information collection was Meta's open-source release of LLama 2 in July 2023 [61] (Figure 3b). Following the release, users discussed the privacy risks of using cloud-based conversational AI platforms and the potential privacy



(a) Personal data (b) Proprietary information (c) Third-party sharing

(d) GDPR compliance (e) Platform security (f) AI-enabled applications

Figure 3: Impact of major events on S&P discussions over time. We only display concerns that are significantly impacted ($p < 0.05$) by at least one event and annotate events that had a significant impact.

benefits of using open-source models. For example, a user noted *"Instead of exposing potentially private code to a public service provider like OpenAI, why not use an open source local LLM (like Llama 2 or StarCoder)"*.

Regulatory and Policy Changes. Italy banned ChatGPT in late March 2023 due to reported GDPR violations [33]. We observe a peak in users' discussions about AI platforms' compliance with GDPR around this event (Figure 3d). The event sparked user discussions around privacy. Some users note their preference for strict privacy laws supporting the ban and GDPR regulations (e.g., *"it's a matter of protection of personal data, how ChatGPT processes the user data is considered 'problematic'"*). Others think that the privacy regulations are *"stupid"* and excessive. For example, one user commented on the ban, saying *"I'm so glad I don't live in Europe. I know we have a lot of bureaucracy in the United States, but it's on a whole [other] level over there."*

Service and Feature Expansions. Our analysis shows that new services and features (e.g., APIs for proprietary models and enterprise plans) influence discussions related to the collection of personal and proprietary information and the security of AI-enabled apps. We observe a peak in concerns related to proprietary information (Figure 3b) and the security of AI-enabled apps (Figure 3f) in March 2023 when OpenAI enabled API access for their proprietary GPT models. Enterprises planning to incorporate these APIs in their operations have concerns about collecting and using sensitive corporate data. Users integrating these APIs in their products and services raise concerns about their susceptibility to exploits. We observe similar trends for proprietary data concerns with ChatGPT Enterprise's launch in late August 2023 [93] and for concerns about the security of AI-enabled apps with the release of Custom GPTs in early November 2023 [86].

We also find that some service expansions impact users' discussions on personal data collection (Figure 3a). For instance, the release of ChatGPT Plus in February 2023 [94] sparked discussions about whether the paid plan would ensure that user data would not be used for model training.

Security Bugs and Breach Reports. We find that security incidents have a direct impact on user discussions related to S&P. In March 2023, a bug in ChatGPT led to users seeing the titles of other users' conversations [18]. This correlated with an increase in concerns about personal data collection (Figure 3a) and platform security (Figure 3e). Users reported incidents like: *"ChatGPT showed me another person's dialogues ... [It] doesn't open [the] chat when clicked, [but] still seems like a serious privacy problem."* Another user questioned the platform's security *"Has ChatGPT or me been hacked? [I've] never had these conversations."*

A similar event occurred in April 2023 when Samsung employees reportedly leaked sensitive company data to ChatGPT [95], leading to the company eventually banning ChatGPT among employees [19]. We find that this event led to increased user discussions about proprietary information (Figure 3b), primarily due to user concerns about the privacy of corporate or company data shared with the AI platform. Enterprises became more cautious, with some banning the use of ChatGPT among employees: *"My company blocked ChatGPT... They are worried about security concerns."*

Key Takeaways (RQ3). Our analysis shows that significant events, such as feature launches, regulatory changes, and security incidents, trigger distinct patterns in users' S&P concerns. Key concerns are data sharing with third parties, corporate data privacy, and platform security, with notable spikes in concern following events like Microsoft's investment in OpenAI and the ChatGPT security bug.

7. Discussion

Our analysis of Reddit posts from the `r/ChatGPT` community provides valuable insights into users' security and privacy (S&P) concerns toward conversational AI platforms. These findings reveal parallels with concerns about other

emerging technologies while highlighting unique challenges posed by conversational AI platforms. In this section, we discuss the implications of our findings, situate them within the broader context of technology adoption and privacy concerns, and provide recommendations for key stakeholders.

7.1. Key Insights and Implications

Although users' concerns and attitudes toward conversational AI platforms mirror those of other computing platforms, such as smartphones and IoT devices, we identify unique factors inherent to conversational AI platforms that escalate or reshape these concerns. Moreover, we find that users' S&P concerns evolve over time, influenced by platform updates, new features, and reported security incidents.

7.1.1. Parallels with Other Technologies. Our study shows how users worry about extensive data collection practices, echoing longstanding concerns in the S&P community. Several of the S&P concerns identified in our analysis parallel those found in other computing platforms and emerging technologies. For instance, concerns about personal data collection, such as PII, location data, and activity tracking, are not unique to AI platforms [96], [97]. Similarly, the unease over data sharing with third parties, the potential misuse of data for targeted advertising, and inadequate transparency and control mechanisms are common themes in S&P research on devices such as smartphones [98] and IoT devices [99] and platforms such as social media [100] and VR [101].

7.1.2. Concerns Unique to Conversational AI Platforms. Despite these similarities, conversational AI platforms present critical differences that exacerbate users' S&P concerns, posing obstacles to secure adoption. First, the conversational format and LLMs' ability to emulate human empathy often lead users to disclose more sensitive or personal information than they would on other platforms (e.g., confidential business plans, proprietary code, or deeply personal struggles).

Second, conversational AI platforms use user data for model training, which can lead to the inadvertent memorization of sensitive information by AI models, a phenomenon distinct from traditional data-storage models. Memorized data can be regurgitated during inference, raising ethical and regulatory questions about a "right to be forgotten".

Lastly, conversational AI has seen an unprecedented and rapid scale of adoption. This widespread popularity has driven its integration into high-stakes sectors such as finance, healthcare, and enterprise services, where sensitive data and critical operations are involved. In these contexts, the natural language interface becomes a gateway to high-stakes data, creating novel security risks like prompt injections or jailbreaks that can expose secrets or allow unauthorized access to sensitive functionality.

7.1.3. Validity of User Concerns. As our analysis is based on user-posted content from Reddit, many concerns reflect users' perceptions of platform behavior. While these concerns are not always technically accurate, they offer valuable insights into user sentiment and potential trust or usability issues. We validate user concerns by analyzing platform documentation, feature updates, privacy policies (e.g., OpenAI's), and existing literature, categorizing them as presently valid concerns, resolved issues, or misconceptions.

Presently Valid Concerns. We find that several concerns raised by users remain pressing and unresolved. For instance, LLM-generated code has been shown to contain security vulnerabilities, such as susceptibility to SQL injection and cross-site scripting attacks [102]. Platforms remain vulnerable to prompt injections and jailbreaks, with studies showing how maliciously crafted inputs can override system instructions or expose sensitive information [57], [103]. Concerns about memorization and lack of "right to be forgotten" also persist; prior work has demonstrated that LLMs can retain and regurgitate sensitive data from training sets, raising regulatory and ethical challenges [16]. Lastly, model training and memory remain opted in by default on ChatGPT [89], which is a pre-selection dark pattern[104].

Resolved Issues. Some concerns reflect actual incidents or bugs that platform developers have since addressed. For example, the ChatGPT bug that briefly exposed conversation titles to other users was acknowledged by OpenAI and is resolved [18]. Similarly, model training and chat history, which were initially coupled together [105] (forced action dark pattern[104]), have now been decoupled [89].

Misconceptions. Concerns about platforms accessing GPS location data without permission are unsubstantiated. We also found no discrepancies in features or access between accounts with and without model training disabled. These misconceptions imply that users have incorrect mental models of S&P and conversational AI platforms' capabilities, indicating the need for platforms to educate users by making privacy-related information more accessible (Section 7.2.2).

7.1.4. Evolution of Users' Concerns. Our findings indicate that users' S&P concerns evolve over time, often in response to new features and reported incidents. Initially, many users express excitement about conversational AI platforms' capabilities but simultaneously exhibit caution due to underlying S&P concerns. We find that the introduction of new functionalities (e.g., custom GPTs), tends to reignite discussions about privacy, primarily due to a lack of transparency and detailed information about these features.

Moreover, users' concerns are exacerbated by the perceived inadequacy of transparency and control mechanisms. Confusion over privacy settings, policy changes, and data control options contributes to a sense of mistrust. Users often feel that their ability to manage their data is limited or obfuscated, leading to frustration and skepticism about the platforms' commitment to protecting their privacy.

Reported security incidents, such as data breaches or software bugs that expose user data, also significantly impact users' trust in the platforms. For example, the ChatGPT bug that revealed conversation titles to other users led to heightened concerns about data security and platform reliability. These incidents underscore the importance of

robust security measures and transparent communication from platform providers to maintain user trust.

7.2. Recommendations

Based on our findings, we synthesize recommendations for stakeholders to address users' S&P concerns and promote the safe, responsible use of conversational AI platforms.

7.2.1. For Users. Users play an active role in protecting their privacy and ensuring safe interactions through informed decisions and cautious behavior.

Exercise Data Controls. Users should explore and use the privacy settings available on the platforms, such as opting out of data usage for model training when possible. Adjusting these settings helps limit the exposure of personal information and aligns with individual privacy preferences.

Sanitize Inputs. To minimize the risk of unintentional data exposure, users should avoid sharing sensitive or personally identifiable information (PII) in prompts. Simple strategies (e.g., replacing names, addresses, or other identifiers with placeholders) can provide an effective first layer of privacy protection. Prior work has demonstrated that prompt-level modifications, such as removing, masking, or substituting sensitive keywords, can substantially reduce privacy risks while preserving the utility of LLM output [106].

Read and Compare Privacy Information. Taking time to review platforms' privacy policies, especially simplified summaries when available, can help users make informed choices. Comparing data practices across platforms enables users to select services that align with their S&P goals.

7.2.2. For Platforms. Conversational AI platforms bear significant responsibility in safeguarding user data and maintaining trust. To address users' concerns, platforms should enhance transparency, provide better control mechanisms, and prioritize user privacy in their operations.

Improve Transparency. Platforms should prioritize clear and accessible communication regarding data handling practices. Users often find privacy policies and terms of service documents dense and difficult to understand, leading to confusion and mistrust. To address this, platforms can take inspiration from the mobile ecosystem, where iOS and Android have introduced privacy labels [107] and data safety sections [108] that simplify key information and make it more user-friendly. By adopting similar design patterns and highlighting data usage, training purposes, and sharing practices, AI platforms can empower users to make informed decisions and compare privacy standards across services, ultimately enhancing trust and transparency.

Design Better Data Controls and Nudges. Platforms can help alleviate concerns by offering robust and user-friendly data controls. Users have expressed frustration over the lack of clear settings to manage their data sharing preferences, especially regarding opting out of model training without losing essential features. Platforms can empower users to take control of their personal information by designing intuitive interfaces that allow them to easily adjust their privacy settings. Gentle interventions and nudges that direct users toward safer practices can help improve users' attitudes towards S&P. Such proactive nudging strategies have shown promise in other domains. For instance, Wang et al. [109] found that privacy nudges on Facebook could reduce unintended disclosures, while Li et al. [39] advocate for privacy-focused nudges during smart home device setup.

Educate Users. Prior research in S&P has shown that even motivated users often struggle to make safe choices without clear, accessible guidance [110]. Building on these insights, we recommend that platforms develop educational resources to help users understand their operations and data-handling practices. These resources (e.g., tutorials, FAQ pages) can address common misconceptions, provide guidance on best practices for protecting personal data (e.g., adjusting privacy settings), and explain how user data is used for model training. By providing clear, targeted education, platforms can foster a more informed user base.

7.2.3. For Enterprises. Organizations using conversational AI in their operations must proactively address S&P concerns to protect both their interests and those of their stakeholders.

Develop Clear Usage Guidelines. Developing clear guidelines for the use of AI platforms is essential. Enterprises should establish policies that outline acceptable practices for employees, particularly regarding the handling of sensitive or proprietary information. For instance, guidelines might prohibit the input of confidential company data into AI platforms unless certain privacy safeguards are in place. Training programs can educate employees about the risks and best practices, helping to prevent potential data leaks.

Regular Security Assessments. Conducting thorough security assessments of AI-powered applications is another critical step. Enterprises should evaluate their systems for vulnerabilities, such as prompt injections or jailbreaks, which could compromise security or allow unauthorized access. Implementing technical safeguards, regular audits, and continuous monitoring can enhance the resilience of these applications against potential threats. Leveraging enterprise solutions offered by AI platforms can also address privacy and compliance requirements. Using these enterprise-grade services, organizations can ensure that their use of AI aligns with legal standards and protects sensitive information.

7.2.4. For Policymakers. Policymakers have a pivotal role in establishing frameworks that protect users while allowing technological innovation.

Establish Clear Regulations. Policymakers should develop standards that define acceptable data practices for AI platforms, including requirements for transparency, user consent, and data protection. These regulations should specifically address issues such as data usage in model training, retention policies, and the handling of PII. For instance, laws could mandate that platforms obtain explicit consent before using user data for training purposes or sharing it with third parties.

Promote Standardization of Privacy Information. By encouraging standardized privacy labels or data safety sections,

policymakers can help users better understand and compare the data practices of different platforms. This standardization can drive industry-wide improvements in privacy standards and empower users to make informed choices.

7.3. Limitations and Future Work

Our study focuses exclusively on the `r/ChatGPT` subreddit, a leading discussion forum for conversational AI with a substantial user base. While this forum provides rich insights into user perspectives, we acknowledge our study's limitations. First, Reddit users tend to be more technologically savvy, and `r/ChatGPT` participants may possess greater knowledge about conversational AI platforms compared to the general public. While `r/ChatGPT` offers a large, active user base with a diverse range of perspectives, including discussions on alternative platforms and open-source models, it may not fully represent the broader population of conversational AI users. Users who frequent Reddit or this specific subreddit may differ demographically or attitudinally from users engaging with conversational AI in other contexts (e.g., enterprise, non-English communities). These factors introduce a selection bias, potentially limiting the applicability of our findings to a broader audience.

Second, Reddit's pseudonymous nature limits access to demographic data such as profession, age, or location, making it difficult to contextualize user concerns or analyze trends across specific demographics. Future work will extend the analysis to other platforms such as Twitter, Discord, and community forums to capture a more diverse range of user perspectives and richer demographic or contextual data.

Third, while our S&P-related posts classifier achieves high accuracy on our seed corpus, there may still be false positives or negatives affecting the reported prevalence of different themes. Automated classification may not fully capture the nuances of user concerns, potentially leading to overestimations or underestimations of certain issues. Moreover, some posts may include users quoting AI-generated content or users presenting AI-generated content as their own. Posts that are clearly AI-generated (e.g., "ChatGPT's views on humanity") or unrelated to S&P are labeled as not-S&P during manual annotation and accordingly filtered by our S&P classifier. Yet, distinguishing AI-generated text from user-written content remains an open challenge [111].

Given that our findings reflect specific concerns from professionals, developers, and other users using these platforms for specific applications, future work could explore users' S&P concerns and attitudes in specific contexts and domains. By examining sector-specific issues, we can identify unique challenges and develop targeted recommendations for users and stakeholders in those fields. Similarly, future research should investigate the influence of geopolitical and cultural factors on user attitudes toward conversational AI platforms to understand regional differences and develop culturally sensitive solutions and policies.

8. Conclusion

We conduct a large-scale analysis of online user posts to study users' S&P concerns and attitudes toward conversational AI platforms. Our qualitative analysis shows that users are concerned with all stages of the data lifecycle—collection, usage, and retention—and seek better security, regulatory compliance, transparency, and control over their data. We also highlight how users' concerns evolve over time in response to major events. Based on our findings, we provide recommendations for different stakeholders involved.

Acknowledgments

We thank our anonymous reviewers and shepherd for providing us with valuable feedback. This work has been partially supported by UCI Academic Senate Council on Research, Computing and Libraries (CORCL) Award. Any findings, conclusions, and recommendations expressed in this paper are those of the authors only.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[2] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann *et al.*, "Palm: Scaling language modeling with pathways," *Journal of Machine Learning Research*, 2023.

[3] ChatGPT. https://chatgpt.com. [Online; accessed: 03-April-2025].

[4] Gemini. https://gemini.google.com. [Online; accessed: 03-April-2025].

[5] Claude. https://claude.ai/. [Online; accessed: 03-April-2025].

[6] K. Pandya and M. Holia, "Automating customer service using langchain: Building custom open-source gpt chatbot for organizations," *arXiv preprint arXiv:2310.05421*, 2023.

[7] Automating Customer Service using LangChain: Building custom open-source GPT Chatbot for organizations. https://www.forbes.com/sites/sunilrajaraman/2024/06/18/ai-driven-customer-service-is-gaining-steam/, 2024. [Online; accessed: 03-April-2025.

[8] Conversational AI Revolutionizes the Customer Experience Landscape. https://www.technologyreview.com/2024/02/26/1088846/conversational-ai-revolutionizes-the-customer-experience-landscape/, 2024. [Online; accessed: 03-April-2025].

[9] Introducing Gemini, your new personal AI assistant. https://gemini.google/assistant/. [Online; accessed: 03-April-2025].

[10] Apple Intelligence. https://www.apple.com/apple-intelligence/. [Online; accessed: 03-April-2025].

[11] The Rise of Large Language Models: A Helping Hand for Healthcare. https://www.forbes.com/councils/forbesbusinesscouncil/2024/05/29/the-rise-of-large-language-models-a-helping-hand-for-healthcare/, 2024. [Online; accessed: 03-April-2025].

[12] J. Clusmann, F. R. Kolbinger, H. S. Muti, Z. I. Carrero, J.-N. Eckardt, N. G. Laleh, C. M. L. Löffler, S.-C. Schwarzkopf, M. Unger, G. P. Veldhuizen *et al.*, "The future landscape of large language models in medicine," *Communications medicine*, 2023.

[13] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann, "Bloomberggpt: A large language model for finance," *arXiv preprint arXiv:2303.17564*, 2023.

[14] Y. Li, S. Wang, H. Ding, and H. Chen, "Large language models in finance: A survey," in *ACM international conference on AI in finance*, 2023.

[15] C. Ischen, T. Araujo, H. Voorveld, G. van Noort, and E. Smit, "Privacy concerns in chatbot interactions," in *Chatbot Research and Design: CONVERSATIONS 2019*. Springer, 2020.

[16] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson *et al.*, "Extracting training data from large language models," in *USENIX Security Symposium*, 2021.

[17] R. Staab, M. Vero, M. Balunovic, and M. Vechev, "Beyond memorization: Violating privacy via inference with large language models," in *International Conference on Learning Representations (ICLR)*, 2024.

[18] OpenAI CEO admits a bug allowed some ChatGPT users to see others' conversation titles. https://www.cnbc.com/2023/03/23/openai-ceo-says-a-bug-allowed-some-chatgpt-to-see-others-chat-titles.html, 2023. [Online; accessed: 03-April-2025].

[19] Samsung Bans ChatGPT and Other Chatbots for 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/, 2023. [Online; accessed: 03-April-2025].

[20] Wall Street Banks Are Cracking Down on AI-powered ChatGPT. https://www.bloomberg.com/news/articles/2023-02-24/citigroup-goldman-sachs-join-chatgpt-crackdown-fn-reports, 2023. [Online; accessed: 03-April-2025].

[21] Amazon Warns Employees Not to Share Confidential Information with ChatGPT. https://www.businessinsider.com/amazon-chatgpt-openai-warns-employees-not-share-confidential-information-microsoft-2023-1, 2023. [Online; accessed: 03-April-2025].

[22] Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence, 2023. [Online; accessed: 03-April-2025].

[23] J. Jang, D. Yoon, S. Yang, S. Cha, M. Lee, L. Logeswaran, and M. Seo, "Knowledge unlearning for mitigating privacy risks in language models," in *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

[24] N. Kandpal, E. Wallace, and C. Raffel, "Deduplicating training data mitigates privacy risks in language models," in *International Conference on Machine Learning (ICML)*, 2022.

[25] D. Yu, S. Naik, A. Backurs, S. Gopi, H. A. Inan, G. Kamath, J. Kulkarni, Y. T. Lee, A. Manoel, L. Wutschitz *et al.*, "Differentially private fine-tuning of language models," in *International Conference on Learning Representations (ICLR)*, 2022.

[26] P. Lison, I. Pilán, D. Sanchez, M. Batet, and L. Øvrelid, "Anonymisation models for text data: State of the art, challenges and future directions," in *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

[27] Z. Zhang, M. Jia, H.-P. Lee, B. Yao, S. Das, A. Lerner, D. Wang, and T. Li, ""It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents," in *CHI Conference on Human Factors in Computing Systems*, 2024.

[28] E. Gumusel, K. Z. Zhou, and M. R. Sanfilippo, "User privacy harms and risks in conversational ai: A proposed framework," *arXiv preprint arXiv:2402.09716*, 2024.

[29] Reddit. https://www.reddit.com. [Online; accessed: 03-April-2025].

[30] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[31] D. McDowall, R. McCleary, and B. J. Bartos, *Interrupted time series analysis*. Oxford University Press, 2019.

[32] Microsoft Confirms Its $10 Billion Investment Into ChatGPT, Changing How Microsoft Competes With Google, Apple, and Other Tech Giants. https://www.forbes.com/sites/qai/2023/01/27/microsoft-confirms-its-10-billion-investment-into-chatgpt-changing-how-microsoft-competes-with-google-apple-and-other-tech-giants/, 2023. [Online; accessed: 03-April-2025].

[33] Italy Blocks ChatGPT Over Privacy Concerns, https://www.nytimes.com/2023/03/31/technology/chatgpt-italy-ban.html, 2023. [Online; accessed: 03-April-2025].

[34] P. Kumaraguru and L. F. Cranor, "Privacy indexes: a survey of westin's studies," 2005.

[35] S. Kokolakis, "Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon," *Computers & security*, 2017.

[36] J. Lin, B. Liu, N. Sadeh, and J. I. Hong, "Modeling {Users'} mobile app privacy preferences: Restoring usability in a sea of permission settings," in *USENIX Symposium On Usable Privacy and Security (SOUPS)*, 2014.

[37] P. E. Naeini, S. Bhagavatula, H. Habib, M. Degeling, L. Bauer, L. F. Cranor, and N. Sadeh, "Privacy expectations and preferences in an iot world," in *USENIX Symposium On Usable Privacy and Security (SOUPS)*, 2017.

[38] J. L. Dupree, R. Devries, D. M. Berry, and E. Lank, "Privacy personas: Clustering users via attitudes and behaviors toward security practices," in *CHI Conference on Human Factors in Computing Systems*, 2016.

[39] J. Li, K. Sun, B. S. Huff, A. M. Bierley, Y. Kim, F. Schaub, and K. Fawaz, ""It's up to the Consumer to be Smart": Understanding the Security and Privacy Attitudes of Smart Home Users on Reddit," in *IEEE Symposium on Security and Privacy (SP)*, 2023.

[40] E. Zeng, S. Mare, and F. Roesner, "End user security and privacy concerns with smart homes," in *USENIX Symposium On Usable Privacy and Security (SOUPS)*, 2017.

[41] J. M. Haney, J. M. Haney, S. M. Furman, and Y. Acar, *User perceptions of smart home privacy and security*. US Department of Commerce, National Institute of Standards and Technology, 2020.

[42] M. Tabassum, T. Kosinski, and H. R. Lipford, ""I don't own the data": End User Perceptions of Smart Home Device Data Practices and Risks," in *USENIX Symposium On Usable Privacy and Security (SOUPS)*, 2019.

[43] E. Zeng and F. Roesner, "Understanding and improving security and privacy in multi-user smart homes: A design exploration and in-home user study," in *USENIX Security Symposium*, 2019.

[44] A. A. Hossain and W. Zhang, "Privacy and security concern of online social networks from user perspective," in *International Conference on Information Systems Security and Privacy (ICISSP)*, 2015.

[45] M. Al-Kfairy, A. Al-Adaileh, M. Tubishat, O. Alfandi, M. BinAmro, and A. Alomari, "A sentiment analysis approach for identifying users' security and privacy perception of metaverse in twitter," in *International Conference on Smart Applications, Communications and Networking (SmartNets)*, 2023.

[46] M. Wei, M. Stamos, S. Veys, N. Reitinger, J. Goodman, M. Herman, D. Filipczuk, B. Weinshel, M. L. Mazurek, and B. Ur, "What twitter knows: Characterizing ad targeting practices, user perceptions, and ad explanations through users' own twitter data," in *USENIX Security Symposium*, 2020.

[47] M. Tahaei, K. Vaniea, and N. Saphra, "Understanding privacy-related questions on stack overflow," in *CHI conference on human factors in computing systems*, 2020.

[48] S. Vetrivel, V. Van Harten, C. H. Gañán, M. Van Eeten, and S. Parkin, "Examining consumer reviews to understand security and privacy issues in the market of smart home devices," in *USENIX Security Symposium*, 2023.

[49] T. Li, E. Louie, L. Dabbish, and J. I. Hong, "How developers talk about personal data and what it means for user privacy: A case study of a developer forum on reddit," *Proceedings of the ACM on Human-Computer Interaction*, 2021.

[50] M. Wei, S. Consolvo, P. G. Kelley, T. Kohno, T. Matthews, S. Meiklejohn, F. Roesner, R. Shelby, K. Thomas, and R. Umbach, "Understanding Help-Seeking and Help-Giving on social media for Image-Based sexual abuse," in *USENIX Security Symposium*, 2024.

[51] S. Horawalavithana, A. Bhattacharjee, R. Liu, N. Choudhury, L. O. Hall, and A. Iamnitchi, "Mentions of security vulnerabilities on reddit, twitter and github," in *IEEE/WIC/ACM International Conference on Web Intelligence*, 2019.

[52] H. Harkous, S. T. Peddinti, R. Khandelwal, A. Srivastava, and N. Taft, "Hark: A deep learning system for navigating privacy feedback at scale," in *IEEE Symposium on Security and Privacy (SP)*, 2022.

[53] O. Akgul, S. T. Peddinti, N. Taft, M. L. Mazurek, H. Harkous, A. Srivastava, and B. Seguin, "A decade of Privacy-Relevant android app reviews: Large scale trends," in *USENIX Security Symposium*, 2024.

[54] D. Mukherjee, A. Ahmadi, M. V. Pour, and J. Reardon, "An empirical study on user reviews targeting mobile apps' security & privacy," *arXiv preprint arXiv:2010.06371*, 2020.

[55] P. Nema, P. Anthonysamy, N. Taft, and S. T. Peddinti, "Analyzing user perspectives on mobile app privacy at scale," in *International Conference on Software Engineering (ICSE)*, 2022.

[56] D. C. Nguyen, E. Derr, M. Backes, and S. Bugiel, "Short text, large effect: Measuring the impact of user reviews on android app security & privacy," in *IEEE Symposium on Security and Privacy (SP)*, 2019.

[57] Y. Liu, Y. Jia, R. Geng, J. Jia, and N. Z. Gong, "Formalizing and benchmarking prompt injection attacks and defenses," in *USENIX Security Symposium*, 2024.

[58] H.-P. H. Lee, L. Gao, S. Yang, J. Forlizzi, and S. Das, ""I Don't Know If We're Doing Good. I Don't Know If We're Doing Bad": Investigating How Practitioners Scope, Motivate, and Conduct Privacy Work When Developing AI Products," in *USENIX Security Symposium*, 2024.

[59] ShareGPT52K. https://huggingface.co/datasets/RyokoAI/ShareGPT52K. [Online; accessed: 03-April-2025].

[60] General Data Protection Regulation (GDPR). https://gdpr-info.eu/. [Online; accessed: 03-April-2025].

[61] Meta Introduces Llama 2: A New Open-Source AI Model, https://about.fb.com/news/2023/07/llama-2/, 2023. [Online; accessed: 03-April-2023].

[62] Pushshift. https://pushshift.io/. [Online; accessed: 03-April-2025.

[63] S. Hurtado, P. Ray, and R. Marculescu, "Bot detection in reddit political discussion," in *International Workshop on Social Sensing*, 2019.

[64] B. Yan, K. Li, M. Xu, Y. Dong, Y. Zhang, Z. Ren, and X. Cheng, "On protecting the data privacy of large language models (llms): A survey," *arXiv preprint arXiv:2403.05156*, 2024.

[65] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (llm) security and privacy: The good, the bad, and the ugly," *High-Confidence Computing*, 2024.

[66] T. Cui, Y. Wang, C. Fu, Y. Xiao, S. Li, X. Deng, Y. Liu, Q. Zhang, Z. Qiu, P. Li *et al.*, "Risk taxonomy, mitigation, and assessment benchmarks of large language model systems," *arXiv preprint arXiv:2401.05778*, 2024.

[67] X. Yang, Z. Wen, W. Qu, Z. Chen, Z. Xiang, B. Chen, and H. Yao, "Memorization and privacy risks in domain-specific large language models," in *ICLR Reliable and Responsible Foundation Models Workshop*, 2024.

[68] N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramer, and C. Zhang, "Quantifying memorization across neural language models," in *International Conference on Learning Representations (ICLR)*, 2024.

[69] V. Hartmann, A. Suri, V. Bindschaedler, D. Evans, S. Tople, and R. West, "Sok: Memorization in general-purpose large language models," *arXiv preprint arXiv:2310.18362*, 2023.

[70] P. Emami-Naeini, J. Dheenadhayalan, Y. Agarwal, and L. F. Cranor, "Which privacy and security attributes most impact consumers' risk perception and willingness to purchase iot devices?" in *IEEE Symposium on Security and Privacy (SP)*, 2021.

[71] P. Emami-Naeini, H. Dixon, Y. Agarwal, and L. F. Cranor, "Exploring how privacy and security factor into iot device purchase behavior," in *CHI Conference on Human Factors in Computing Systems*, 2019.

[72] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention," in *International Conference on Learning Representations (ICLR)*, 2021.

[73] GPT-4o. https://platform.openai.com/docs/models/gpt-4o. [Online; accessed: 03-April-2025].

[74] Gemini models. https://ai.google.dev/gemini-api/docs/models, [Online; accessed: 03-April-2025].

[75] B. Saunders, J. Sim, T. Kingstone, S. Baker, J. Waterfield, B. Bartlam, H. Burroughs, and C. Jinks, "Saturation in qualitative research: exploring its conceptualization and operationalization," *Quality & quantity*, 2018.

[76] N. McDonald, S. Schoenebeck, and A. Forte, "Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice," *Proceedings of the ACM on human-computer interaction*, 2019.

[77] Y. Liu, Y. Lu, H. Liu, Y. An, Z. Xu, Z. Yao, B. Zhang, Z. Xiong, and C. Gui, "Hierarchical prompt learning for multi-task learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[78] Guidance on the Protection of Personal Identifiable Information. https://www.dol.gov/general/ppii. [Online; accessed: 03-April-2025].

[79] Google Workspace. https://workspace.google.com/solutions/ai/. [Online; accessed: 03-April-2025].

[80] ChatGPT Pricing. https://openai.com/chatgpt/pricing/. [Online; accessed: 03-April-2025].

[81] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[82] A. Juels, "Targeted advertising... and privacy too," in *Cryptographers' Track at the RSA Conference*. Springer, 2001.

[83] OpenAI Privacy Policy. https://openai.com/policies/row-privacy-policy/, 2024. [Online; accessed: 03-April-2025].

[84] H. Li, D. Guo, W. Fan, M. Xu, J. Huang, F. Meng, and Y. Song, "Multi-step jailbreaking privacy attacks on chatgpt," in *EMNLP*, 2023.

[85] F. Perez and I. Ribeiro, "Ignore previous prompt: Attack techniques for language models," in *NeurIPS ML Safety Workshop*, 2022.

[86] Introducing GPTs. https://openai.com/index/, 2023. [Online; accessed: 03-April-2025].

[87] Introducing the GPT Store. https://openai.com/index/introducing-the-gpt-store/, 2023. [Online; accessed: 03-April-2025].

[88] Health Insurance Portability and Accountability Act of 1996 (HIPAA). https://www.cdc.gov/phlp/php/resources/health-insurance-portability-and-accountability-act-of-1996-hipaa.html. [Online; accessed: 03-April-2025].

[89] Data Controls FAQ. https://help.openai.com/en/articles/7730893-data-controls-faq. [Online; accessed: 03-April-2025].

[90] Llama. https://www.llama.com/. [Online; accessed: 03-April-2025].

[91] OpenAI, "Memory and new controls for chatgpt," 2023, accessed: 2024-11-13. [Online]. Available: https://openai.com/index/memory-and-new-controls-for-chatgpt/

[92] Y. Hochberg and A. C. Tamhane, *Multiple Comparison Procedures*. John Wiley & Sons, 1987.

[93] Introducing ChatGPT Enterprise. https://openai.com/index/introducing-chatgpt-enterprise/, 2023. [Online; accessed: 03-April-2025].

[94] Introducing ChatGPT Plus. https://openai.com/index/chatgpt-plus/, 2023. [Online; accessed: 03-April-2025].

[95] Samsung employees leaked corporate data in ChatGPT, https://www.ciodive.com/news/Samsung-Electronics-ChatGPT-leak-data-privacy/647137/, 2023. [Online; accessed: 03-April-2025].

[96] L. Agarwal, N. Shrivastava, S. Jaiswal, and S. Panjwani, "Do not embarrass: Re-examining user concerns for online tracking and advertising," in *USENIX Symposium On Usable Privacy and Security (SOUPS)*, 2013.

[97] J. S. Seberger and S. Patil, "Us and them (and it): Social orientation, privacy concerns, and expected use of pandemic-tracking apps in the united states," in *CHI Conference on Human Factors in Computing Systems*, 2021.

[98] A. Frik, J. Kim, J. R. Sanchez, and J. Ma, "Users' expectations about and use of smartphone privacy and security settings," in *CHI Conference on Human Factors in Computing Systems*, 2022.

[99] S. Zheng, N. Apthorpe, M. Chetty, and N. Feamster, "User perceptions of smart home IoT privacy," *Proceedings of the ACM on human-computer interaction*, 2018.

[100] E. Rader, "Awareness of behavioral tracking and information privacy concern in facebook and google," in *USENIX Symposium On Usable Privacy and Security (SOUPS)*, 2014.

[101] S. Abhinaya, A. Agrawal, Y. Yao, Y. Zou, and A. Das, ""What are they gonna do with my data?": Privacy Expectations, Concerns, and Behaviors in Virtual Reality," *Privacy Enhancing Technologies (PETs)*, 2025.

[102] H. Pearce, B. Ahmad, B. Tan, B. Dolan-Gavitt, and R. Karri, "Asleep at the Keyboard? Assessing the Security of GitHub Copilot's Code Contributions," in *IEEE Symposium on Security and Privacy (SP)*, 2022.

[103] ChatGPT's New Code Interpreter Has Giant Security Hole, Allows Hackers to Steal Your Data. https://www.tomshardware.com/news/chatgpt-code-interpreter-security-hole, 2023. [Online; accessed: 03-April-2025].

[104] Z. Shi, R. Sun, J. Chen, J. Sun, M. Xue, Y. Gao, F. Liu, and X. Yuan, "50 shades of deceptive patterns: A unified taxonomy, multimodal detection, and security implications," in *International World Wide Web Conference (WWW)*, 2025.

[105] New ways to manage your data in ChatGPT. https://openai.com/index/new-ways-to-manage-your-data-in-chatgpt/, 2023. [Online; accessed: 03-April-2025].

[106] Y. Zhu, N. Gao, X. Liang, and H. Zhang, "Exploiting privacy preserving prompt techniques for online large language model usage," in *IEEE Global Communications Conference (GLOBECOM)*, 2024.

[107] App privacy details on the App Store. https://developer.apple.com/app-store/app-privacy-details/. [Online; accessed: 03-April-2025].

[108] Data Safety. https://developer.android.com/google/play/integrity/other. [Online; accessed: 03-April-2025].

[109] Y. Wang, P. G. Leon, K. Scott, X. Chen, A. Acquisti, and L. F. Cranor, "Privacy nudges for social media: an exploratory facebook study," in *International World Wide Web Conference (WWW)*, 2013.

[110] A. Whitten and J. Tygar, "Why johnny can't encrypt," *USENIX Security Symposium*, 2005.

[111] J. Wu, R. Zhan, D. Wong, S. Yang, X. Yang, Y. Yuan, and L. Chao, "Detectrl: Benchmarking llm-generated text detection in real-world scenarios," *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Appendix A. Subreddits Related to Conversational AI

Table 1 shows the list of popular subreddits related to conversational AI. We explored these subreddits through manual review and keyword searches and found concerns and attitudes consistent with those observed in `r/ChatGPT`.

TABLE 1: Top subreddits related to conversational AI, ranked by size.

Subreddit	Members
r/ChatGPT	7.4M
r/OpenAI	1.9M
r/CharacterAI	1.7M
r/ChatGPTPro	271K
r/LocalLLaMA	226K
r/ChatGPTCoding	147K
r/ClaudeAI	70K
r/Bard	44K

Appendix B. Keywords for Seed Corpus Creation

Table 3 shows the list of keywords for filtering posts.

Appendix C. S&P Classification Results

Table 2 shows the performance of different models on S&P classification task.

TABLE 2: S&P Post classification results.

Classifier	Accuracy	Precision	Recall	F1 Score
DeBERTa[72]	0.97	0.91	0.76	0.83
RoBERTa [30]	0.96	0.83	0.81	0.82
GPT-4o [73]	0.91	0.71	0.92	0.80
Gemini-1.5-Flash [74]	0.76	0.45	0.95	0.61

Appendix D. S&P-Related Posts Over Time

Figure 5 shows the weekly volume of S&P-related posts over time, averaging 344 posts per week.

Appendix E. Multi-Class Classification Results

Table 4 shows the performance of the multi-class classification of S&P-related posts using GPT-4o [73].

Appendix F. Events Related to Conversational AI Ecosystem

Figure 4 shows the timeline of major events related to conversational AI platforms.

TABLE 3: Keywords used to filter candidate posts for manual annotation. We applied regex-based matching to capture phrase variations (e.g., `r'delete(?:my|your)?data'` to match "delete my data" and "delete your data"). Certain terms like `access` were matched as standalone words to avoid unrelated results (e.g., `accessibility`).

Group	Description	Keywords
General Keywords	Broad S&P-related concerns.	privacy, security, permission, encryption, malicious, steal, access, secure, safe, confidential
Sensitive and Personal Data	Types of sensitive information users worry about.	personal/private/confidential/sensitive/corporate/company/client/customer/health/medical/patient + info/data/doc/file, pii, personally identifiable info
Data Protection Laws	Legal or regulatory references.	gdpr, ccpa, hipaa, eu ai act, data regulation, data compliance, data protection
Storage and Deletion	Data deletion, retention, and storage practices.	data/account/chat/conversation + deletion/removal, data storage, data retention, delete + chat/conversation/account/data
Collection and Usage	How data is gathered, used, or shared.	data collection, surveillance, monitoring, data usage, data handling, data sharing, data selling, opt-out, privacy settings, data controls, disable/turn off + history/memory/training
Security Threats	System vulnerabilities and attacks.	prompt leak, prompt injection, jailbreak, guardrail, model attack, data theft, data + breach/leak/exposure/extraction

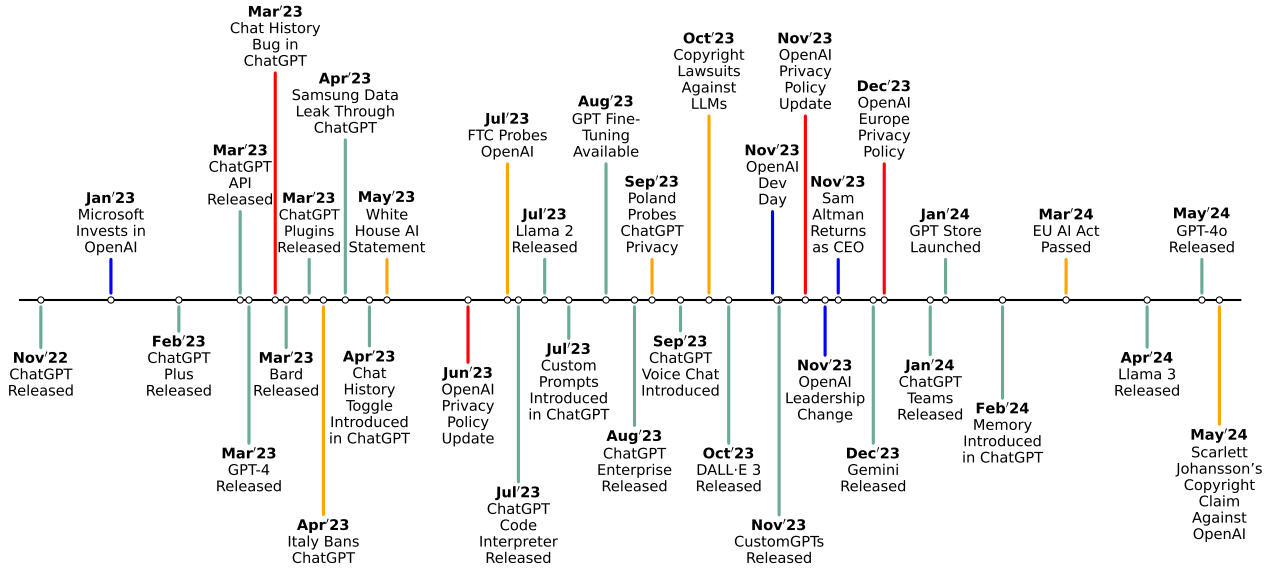


Figure 4: Key events in the development and adoption of conversational AI platforms, including product and feature releases (green), policy and regulatory changes (yellow), corporate developments (navy), and privacy and security events (red).

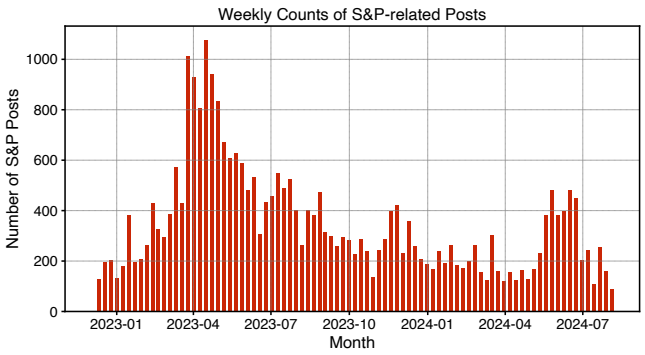


Figure 5: Weekly counts of S&P-related posts.

TABLE 4: Performance of multi-class classification.

Type of Concern	Accuracy	Precision	Recall	F1 Score
Data Collection	0.94	1.0	0.89	0.94
Personal Data	0.97	1.0	0.94	0.97
Proprietary Information	0.96	1.0	0.81	0.90
Data Usage	0.94	0.94	0.91	0.92
Model Training	0.95	0.93	0.82	0.87
Third-party Sharing	0.96	1.0	0.85	0.92
Data Retention	0.97	0.83	1.0	0.91
Security Vulnerabilities	0.96	0.87	0.93	0.90
Platform Security	0.96	0.78	0.88	0.82
LLM-powered Apps	1.0	1.0	1.0	1.0
LLM-generated Code	0.99	0.8	1.0	0.89
Legal Compliance	0.99	1.0	0.88	0.93
GDPR	0.99	1.0	0.8	0.89
HIPAA	1.0	1.0	1.0	1.0
Transparency & Control	0.95	0.71	1.0	0.83

Appendix G.
Meta-Review

The following meta-review was prepared by the program committee for the 2025 IEEE Symposium on Security and Privacy (S&P) as part of the review process as detailed in the call for papers.

G.1. Summary

The paper explores users' security and privacy concerns and attitudes about conversational AI platforms through an analysis of 2.5 million Reddit posts on r/ChatGPT and concludes with recommendations for AI developers and policy makers.

G.2. Scientific Contribution

- Provides a Valuable Step Forward in an Established Field
- Independent Confirmation of Important Results with Limited Prior Research

G.3. Reasons for Acceptance

1) This paper provides a valuable step forward in an established field of concerns relevant to conversational AI by analyzing a large dataset of user concerns about S&P issues, offering a different perspective than surveys or interviews.
2) This paper provides independent confirmation of concerns voiced by users of conversational AI platforms, with the findings highlighting specific challenges and informing recommendations grounded in the data.