Understanding Privacy Norms Around LLM-Based Chatbots: A Contextual Integrity Perspective

Sarah Tran[1], Hongfan Lu[1], Isaac Slaughter[1], Bernease Herman[1], Aayushi Dangol[1], Yue Fu[1], Lufei Chen[1], Biniyam Gebreyohannes[1], Bill Howe[1], Alexis Hiniker[1], Nicholas Weber[1*], Robert Wolfe[2*†]

[1]University of Washington
[2]Rutgers University
saraht45[at]uw.edu, nmweber[at]uw.edu, robert.wolfe[at]rutgers.edu

Abstract

LLM-driven chatbots like ChatGPT have created large volumes of conversational data, but little is known about how user privacy expectations are evolving with this technology. We conduct a survey experiment with 300 US ChatGPT users to understand emerging privacy norms for sharing chatbot data. Our findings reveal a stark disconnect between user concerns and behavior: 82% of respondents rated chatbot conversations as sensitive or highly sensitive — more than email or social media posts — but nearly half reported discussing health topics and over one-third discussed personal finances with ChatGPT. Participants expressed strong privacy concerns ($t(299) = 8.5$, $p < .01$) and doubted their conversations would remain private ($t(299) = -6.9$, $p < .01$). Despite this, respondents uniformly rejected sharing personal data (search history, emails, device access) for improved services, even in exchange for premium features worth $200. To identify which factors influence appropriate chatbot data sharing, we presented participants with factorial vignettes manipulating seven contextual factors. Linear mixed models revealed that only the transmission factors such as informed consent, data anonymization, or the removal of personally identifiable information, significantly affected perceptions of appropriateness and concern for data access. Surprisingly, contextual factors including the recipient of the data (hospital vs. tech company), purpose (research vs. advertising), type of content, and geographic location did not show significant effects. Our results suggest that users apply consistent baseline privacy expectations to chatbot data, prioritizing procedural safeguards over recipient trustworthiness. This has important implications for emerging agentic AI systems that assume user willingness to integrate personal data across platforms.

Introduction

Conversational AI systems like ChatGPT have rapidly become a source of user-generated data, with over 400 million weekly active users creating billions of chat interactions (Beaty 2025). This conversational data represents a fundamentally new category of personal information — more intimate than search queries, more extensive than social media posts, and often containing highly sensitive disclosures about health, finances, and personal relationships. The commercial and strategic value of chatlog data spans targeted advertising (Matz et al. 2024), personalized interfaces (Chen et al. 2024), training next-generation language models (Nasr et al. 2023, 2025), and even geopolitical leverage when nations control chat data from other countries' citizens (Mok 2025; Burgess and Newman 2025). Yet users remain uncertain about how their conversational data is handled. While chatbot providers may offer privacy protections through terms of service agreements, research indicates most users neither understand these policies nor trust that their data will be used responsibly. A 2023 Pew Research Center study found that large majorities of Americans familiar with AI worry about personal information being used in "unintended ways and ways people are not comfortable with" (Faverio and Tyson 2023). These concerns have proved prescient: In December 2024, the chatbot provider WotNot inadvertently exposed over 340,000 private chat logs through an unprotected cloud system (Arntz 2024), and in August 2025, researchers discovered that Google was indexing and making discoverable ChatGPT conversations that users had inadvertently shared - exposing passport numbers, medical records, and employment histories (Cox 2025).

Despite the apparent risks, prior research finds that users of AI chatbots often disclose sensitive information in chat logs, sometimes relying on flawed mental models of how data will be processed and retained by a model's provider (Zhang et al. 2024). Indeed, the presence of user social security numbers and other personally identifiable information (PII) resulted in the withdrawal of ShareGPT, an open dataset of interactions with ChatGPT used in prior privacy studies (Zhang et al. 2024). The apparent gap between many users' behavior when interacting with chatbots (*e.g.,* sharing sensitive PII) and the public's expressed concern with the privacy of chat data motivates a systematic investigation of privacy norms surrounding user chatbot data. Understanding such norms can help inform the design of chat interfaces as well as policy-level considerations for chat data.

The present work thus investigates chatbot privacy norms using the *contextual integrity* framework, which posits that privacy norms exist in the context of *information flows*. In short, an information exchange can be reduced to a set of transmission principles that govern the exchange of a cer-

*These authors contributed equally.
†Work performed while at the University of Washington

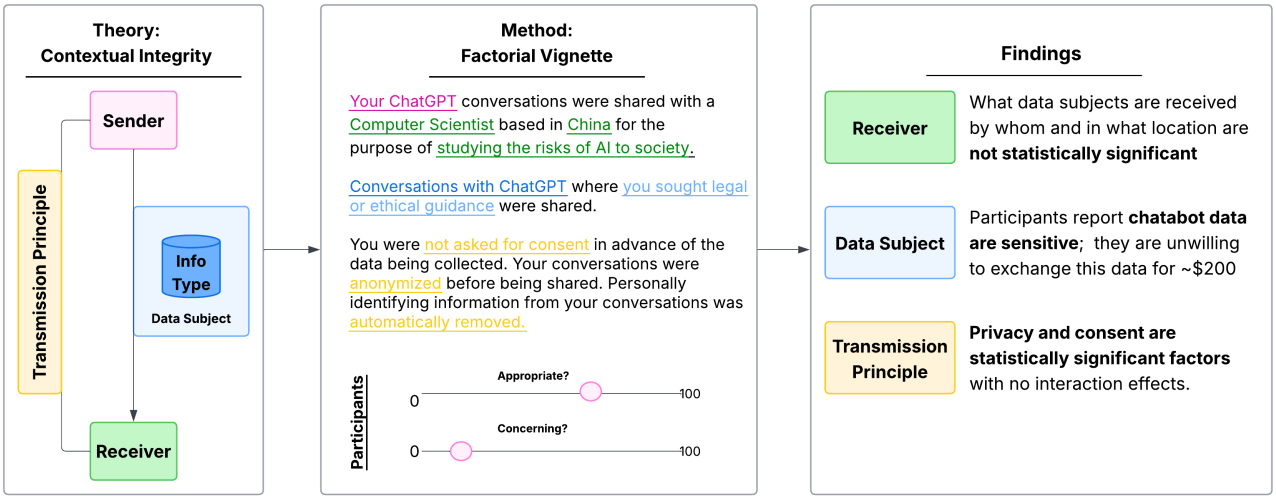

Figure 1: We used the contextual integrity framework to construct factorial vignettes rated by participants based on whether they were appropriate or concerning. This approach allowed us to isolate factors associated with the **transmission principle** - the set of norms and expectations surrounding data exchange - as the primary determinants of privacy norms for chatbot data.

tain kind of information (content) with a certain entity (role) for a certain reason (purpose) - collectively these influence whether the exchange of data is appropriate, or a cause for concern (Nissenbaum 2004; Malkin 2022). Drawing on this theoretical framework, we set out to answer three research questions:

1. **RQ1:** How do users perceive the sensitivity of chatbot data, and how does this compare to their perception of other forms of sensitive personal information?
2. **RQ2:** What factors (*e.g.,* content, purpose, and consent, drawing on a contextual integrity framework) influence users' concerns about the handling of their chatbot data?
3. **RQ3:** How willing are users to integrate chatbot data with data from their device, browser, and social accounts to improve their online experience - either via personalization or by gaining access to premium features?

To address these questions, we drew on methods introduced in past work that addressed the contextual integrity of social media information flows (Gilbert, Vitak, and Shilton 2021), deploying a large-scale survey experiment with N=300 ChatGPT users residing in the U.S. In addition to a battery of questions that directly probe privacy attitudes, each respondent was also presented with 30 vignettes, short stories that each include a randomly drawn level from seven factors of interest, as illustrated in Figure 1. These vignettes allowed us to isolate the factors that influence the two dependent variables rated by respondents: the *appropriateness* of the information flow, and the level of *concern* associated with it. Based on the results of our data analysis, we make three contributions:

1. **Most respondents (82%) consider chatbot conversations sensitive data, and on average agree that they have privacy concerns about ChatGPT conversations.** Respondents also disagreed on average that their conversations with chatbots like ChatGPT would remain private. Furthermore, respondents agreed on average that they were concerned about people receiving inaccurate information from AI; not understanding what AI can do; and having personal information misused by AI.
2. **Only *Informed Consent*, *Anonymity*, and *Privacy* (elements of a transmission principle) influence appropriateness perceptions.** Fitting linear mixed models (LMMs) to vignette ratings reveals that only providing informed consent, anonymizing data, and removing personal information improve respondents' perceptions of the appropriateness of the information flows described by our vignettes. The Role, Purpose, Content Type, and Location factors had no detectable effect on respondent perceptions - possibly because participants view data sharing as problematic regardless of recipient or purpose, or because the meaningful distinctions for users differ from those foregrounded by our experimental design.
3. **Respondents uniformly *reject* integrating personal data with chatbot data to improve their online experience.** Respondents on average disagreed that they would be willing to provide chatbots access to their search history, email, or device in exchange for either 1) better and more personalized chatbot features or 2) a premium-tier ($200/month) chatbot subscription. Similarly, respondents prefer *not* to have their chatbot data used to improve their social media experiences.

Notably, despite their lack of confidence in the ultimate privacy of their chat data, many respondents reported discussing highly sensitive subjects with ChatGPT, including health and wellness (47.0%), personal finances (35.3%), and legal guidance (12.7%), reflecting again the gap between expressions of concern and actual behavior observed in prior work (Zhang et al. 2024; Ngong et al. 2025).

Our findings have implications for AI policy and design.

The primacy of a transmission protocol among our factors suggests the public would benefit from a systematic policy approach that establishes procedures around consent, anonymization, and handling of PII in chatbot data. Moreover, our third contribution suggests that developers may need to reconsider user willingness to exchange access to personal data for more advanced AI features, a key assumption for emerging agentic AI frameworks (Chan et al. 2023, 2024).

Related Work

We first review the related work on the privacy-related attitudes and behaviors of users with respect to their chatbot data, and then introduce the contextual integrity framework.

Chatbot Privacy Attitudes and Concerns

Building on a large body of research that seeks to understand the public's concerns about data privacy and emerging technologies like AI (Golda et al. 2024), recent work has applied both qualitative approaches and methods from computational social science to study users' attitudes toward the data they exchange with chatbots. Alkamli and Alabduljabbar (2024) collect and analyze Twitter data, finding that ChatGPT users express concern about 1) unauthorized access to their chat data and 2) their personal information being shared and exploited across platforms. Ali, Arunasalam, and Farrukh (2025) analyze 2.5 million posts from the r/ChatGPT Reddit community, demonstrating ChatGPT users' concern about the collection and sharing of their personal data. Consistent with broader surveys of the public that demonstrate increases in concern about AI data privacy over time (Faverio 2023; Faverio and Tyson 2023), Ali, Arunasalam, and Farrukh (2025) found increases in the privacy awareness and sensitivity of ChatGPT users in response to reported security incidents and the rollout of new features with questionable effects on user privacy (Ali, Arunasalam, and Farrukh 2025). Recent interview studies similarly find that users express significant concerns about their ability to permanently delete their chat data (Ma et al. 2025), and that they worry about the possibility of chat data being sold to data brokers or shared with third parties (Zhang et al. 2024). Concerns about the collection and secondary use of ostensibly private chat data have also motivated many organizations in domains ranging from medicine (Riedemann, Labonne, and Gilbert 2024) to journalism and fact-checking (Wolfe and Mitra 2024a,b) to adopt the use of *open* chatbot models (Palmer, Smith, and Spirling 2024; Paris, Moon, and Guo 2025; Solaiman 2023), rather than risking the security of client and patient data, or their own valuable proprietary information.

We build on prior work by evaluating how chatbot users' stated privacy preferences vary across contexts. Establishing privacy norms that describe *specific* data collected and exchanged in specific ways is essential for developing effective regulation for general-purpose AI applications like chatbots, which can be deployed across a wide variety of contexts and are known to ingest many forms of sensitive user data. **Our study thus examines the contextual factors that influence privacy expectations using an experimental survey method**, an approach that can isolate specific factors that influence user expectations for privacy.

Chatbot User Privacy Behaviors

Though studies of privacy attitudes find that users express uncertainty about the security of their chat data, research probing users' real-world *behaviors* shows that users nonetheless disclose highly sensitive information to chatbots. Analysis of large, open-source collections of chatbot interactions (such as WildChat (Zhao et al. 2024), LMSYS-Chat-1M (Zheng et al. 2023) or ShareGPT (RyokoAI 2023)) find that users frequently discuss sensitive topics and share PII including email addresses, health conditions, passport numbers, and location data (Mireshghallah et al. 2024). More targeted user studies report that even "privacy conscious" participants often indirectly disclose sensitive information (Ngong et al. 2025). Such disclosures have even prompted the development of privacy-preserving approaches to study the interactions that users engage in with chatbots, such as Anthropic's Claude insights and observations (Clio) platform (Tamkin et al. 2024), which uses a layer of AI assistants to indirectly observe patterns in user data.

Prior work establishes two foundations on which we build: 1. Users' stated privacy preferences clash with their revealed preferences when interacting with chatbots; and, 2. Indirect disclosure of personal information is common across tasks, chatbot providers, and user intent. **Our study elicits users' stated preferences for privacy directly, as they rate the sensitivity of chatbot data against other data sources (*e.g.*, email), and indirectly, by probing their willingness to integrate chatbot data across contexts.**

Contextual Integrity

Contextual Integrity (CI) is both a theory and a diagnostic framework for understanding privacy in a networked information environment (Nissenbaum 2019). As a theory, Nissenbaum and colleagues use CI to describe privacy as the emergent property of norms and practices that govern the appropriate exchange (or flow) of information (Barth et al. 2006). As a diagnostic framework, CI identifies five parameters in an information exchange: data type (the kind of information shared), data subject (what the information is about), sender (who is sharing the data), recipient (who receives the data), and transmission principle (the constraints or norms governing the flow of information) (Malkin 2022; Kumar, Zimmer, and Vitak 2024). CI provides an ontology for making sense of the disruption of expectations when emerging technologies are adopted en masse (Vitak and Zimmer 2023; Apthorpe et al. 2018). Unsurprisingly, then, CI has been used in studies related to consumer facing AI (Mussgnug 2024; Cheng et al. 2024; Ghalebikesabi et al. 2024; Fan et al. 2024; Li et al. 2024), including studies describing indirect disclosures in chatbot interactions (Ngong et al. 2025), and asymmetric privacy expectations between custom GPT creators and users (Ma et al. 2025). Building on prior work using factorial vignettes to understand CI norms (Martin and Nissenbaum 2016), Mireshghallah et al. (2023) used CI-based vignettes to demonstrate that ChatGPT discloses private information in contexts where humans would

not, highlighting a disconnect between the privacy reasoning of large language models and human privacy expectations.

We build on the use of CI in chatbot privacy research in two ways: 1. We deploy a factorial vignette survey based on CI's five parameters, focusing on exchanges of chatbot data; 2. We investigate the effectiveness of privacy interventions - anonymity, confidentiality, consent, and exchange (where a subject receives something in return) - as a transmission principle for chatbot data. **Our study yields evidence for a CI-based approach to chatbot privacy norms, with specific attention paid to CI's transmission principle.**

Approach

We describe the development and deployment of the survey instrument used to address our research questions. We then characterize our participants and our approach to analyzing the survey data we collected. The Institutional Review Board (IRB) at our University reviewed and approved this research.

Survey Instrument and Data Analysis

We used the Qualtrics survey design software to create a five-part survey, described in detail below.

Part 1: Privacy Attitudes. We first asked participants five questions about their attitudes towards general information privacy topics (*e.g.,* "I am concerned that online companies are collecting too much information about me") and about ChatGPT in particular (*e.g.,* "I have privacy concerns about my conversations with ChatGPT"). We also asked respondents to rate their level of concern about the public 1) getting inaccurate information from AI; 2) not understanding what AI can do; and 3) having personal information misused by AI. For general privacy concerns we adopt questions asked in Gilbert, Vitak, and Shilton (2021), and for AI concerns among the public we draw questions from Faverio and Tyson (2023), allowing comparison between studies. Participants responded on a 100-point scale where 0 indicated strong disagreement, 50 neutral, and 100 strong agreement.

Data Analysis. For each question, we report the mean and standard deviation and use a one-sample t-test of significance against a hypothesized mean of 50 (neutral).

Part 2: Private Data Exchange Value. Next, we asked participants about their willingness to exchange access to their search history, emails, device data, or chatbot conversations for 1) personalized chatbot responses and insights; and 2) a premium chatbot subscription valued at $200 (analogous to ChatGPT Pro (OpenAI 2024)). We then asked participants if they would provide access to their chatbot history to improve services like search results, product recommendations, social media post popularity, and social media feed moderation. We used the same 100-point scale as Part 1.

Data Analysis. For each question, we report the mean and standard deviation, and we use a one-sample t-test against a hypothesized mean of 50 (neutral).

Part 3: Factorial Vignettes. Next, participants were shown paragraph-length scenarios (vignettes) describing the use of their ChatGPT conversation history. Each vignette followed the format described in the header of Table 1, with the seven factors denoted in <> tags each randomly replaced with a level included in the table. We drew the Role, Purpose, Content, and Consent/Awareness factors from prior work using factorial vignettes to describe the contextual integrity of information flows (Martin 2012; Gilbert, Vitak, and Shilton 2021). We added the Location factor to account for the newly salient geopolitical dimension of AI data privacy (Mok 2025; Act 2024), and we added the Anonymity and Privacy factors because much recent work notes that many users exchange sensitive and personally identifying data during chatbot conversations (Belen Saglam, Nurse, and Hodges 2021). In total, each participant judged 30 vignettes on two dimensions: "This is an **appropriate** use of my ChatGPT data" (dependent variable 1); and 2. "This use of my data would **concern** me" (dependent variable 2). In total, the N=300 participants in our survey each judged 30 vignettes, yielding 9000 judgments about the use of chatbot data on two dimensions - Appropriateness and Concern.

Data Analysis. We analyze vignette responses by fitting two linear mixed models (LMMs) with Appropriateness (dependent variable 1) and Concern (dependent variable 2). Each LMM is fit to the seven factors described in Table 1, along with a participant random effect. We set the baseline level for the factor *Role* to be "a hospital," for *Purpose* to be "improving user experience with AI," for *Location* to be "the US," for *Content* to be "all conversations with ChatGPT" and for *Consent* to be "asked for consent in advance of data collection." Baseline levels were automatically selected for remaining factors by the LMER package in R, as these factors were either binary or had no clear baseline. Because our models failed assumptions for the homoscedasticity of residuals based on Kolmogorov-Smirnov tests, we used a cluster-robust covariance matrix with CR1 estimate to adjust for our data's multilevel structure (Cameron, Miller et al. 2010).

In addition to the base LMMs, we also compute interaction effects between the five privacy attitude questions (Survey part 1) with 1) the Consent factor and 2) the Location factor. We again fit separate LMMs to the Appropriateness and Concern dependent variables, examining only interactions with Consent and Location, rather than all factors.

Part 4: ChatGPT Usage. Next, we ask participants about their ChatGPT usage, including their account type, date of first use, frequency of use, tasks they typically complete, and topics they frequently discuss. For the last two questions, respondents could select multiple options.

Data Analysis. We report summary statistics as a percentage of participant responses for each question.

Part 5: Chat Data Sensitivity. Finally, we ask respondents to rank the sensitivity of 14 discrete forms of personal information, including their social security number, phone calls, emails, and conversations with chatbots like ChatGPT. Respondents were asked to classify each form of information as "Highly Sensitive," "Sensitive," or "Not Sensitive."

Data Analysis. We use a Friedman Test to check for differences in the sensitivity of the forms of personal information. We then use Wilcoxon signed-rank tests with Bonferroni correction to test for differences between chatbot conversation sensitivity and sensitivity of other forms of information.

Vignette Format: Your ChatGPT conversations were shared with a(n) <Role> based in <Location> for the purpose of <Purpose>. <Content> were shared. You were <Consent/Awareness>. Your conversations were <Anonymity> before being shared. Personally identifying information from your conversations was <Privacy>.	
Factor	**Levels**
Role	A big tech company; A hospital; A government agency; An insurance company; A university computer science researcher; A university social science researcher; A charitable foundation
Location	the United States; the European Union; China
Purpose	training future AI models; creating a public dataset for AI research; improving user experience with AI; fighting terrorism; assessing mental health; personalizing advertising; predicting human behavior; studying the risks of AI to society
Content	all of your conversations with ChatGPT; those conversations with ChatGPT where you used the model to help with your job; those conversations with ChatGPT about your social life and personal relationships; those conversations with ChatGPT about your personal health and wellness; those conversations with ChatGPT where you sought legal or ethical guidance
Consent/Awareness	asked for consent in advance of the data being collected; informed that your data was collected; not be informed that your data was collected
Anonymity	anonymized; not anonymized
Privacy	automatically removed; not removed

Table 1: The format and factors of the factorial vignettes presented in part 3 of our survey instrument. Each vignette is constructed by randomly selecting a level from each of seven factors, and each participant is presented with a total of 30 vignettes.

Gender		**Education**		**Age**		**Politics**		**Race**	
Man	54.0%	HS or less	10.7%	Avg	36	Democrat	48.7%	White	62.3%
Woman	43.3%	Associate's	11.0%	SD	10.8	Republican	23.0%	Black	16.0%
Non-Binary	0.7%	Some college	22.7%	Min	18	Independent	20.3%	Asian	9.0%
Something Else	1.3%	Bachelor's	42.7%	Max	76	Something else	5.7%	Pacific Islander	0.6%
Prefer not say	0.7%	Master's	9.3%			Prefer not say	2.3%	Indigenous	2.0%
		Professional	1.7%					Something Else	3.7%
		Doctorate	1.7%					Other	3.3%
		Prefer not say	0.3%					Prefer not say	1.0%

Table 2: Participant demographics for the N=300 respondents who completed our full survey instrument. Our population overrepresented Men, Democrats, Middle-Aged individuals, and College-Educated individuals relative to the U.S. nationally.

Participants

We recruited survey respondents via CloudResearch Connect (Hartman et al. 2023). Our inclusion criteria required participants to use ChatGPT monthly, be 18 years or older, and reside in the United States. Table 3 describes ChatGPT use among our sample. Most respondents were frequent ChatGPT users, with 46.9% using the application at least weekly and 28.6% using it at least daily. We detail respondent demographic information in the Appendix.

Participants were required to answer all questions in the survey. The mean time to complete the survey was 21 minutes, 33 seconds (SD of 12 minutes, 45 seconds), and the median was 17 minutes, 30 seconds. We manually inspected survey responses completed in less than 10 minutes and discarded those completed in less than 9 minutes.

Preregistration

We submitted a pre-registration of our hypotheses and planned data analysis with the Open Science Foundation in April 2025. This paper is consistent with that preregistration, but we made several significant changes to our modeling and hypothesis testing that prompted us to submit a Statement of Transparent Changes, following best practices in research reporting (Nosek et al. 2018; Lakens 2024). The original pre-registration and the Statement of Transparent Changes can be view here: https://osf.io/m6jt5/

Results

Privacy Attitudes

As seen in Table 4, our results unambiguously show that participants feel concern about the privacy of their chat data and online data more broadly, as they agreed with questions stating 1) that online companies are collecting too much personal information and 2) that they have privacy concerns about their conversations with ChatGPT. Conversely, participants disagreed with questions stating 1) that they trust websites, and 2) that they believe their conversations with chatbots like ChatGPT will remain private. Based on

Year Started Using		Usage Frequency		Account Type	
2023 or earlier	36.9%	Daily	28.6%	No Account	16.2%
2024	57.6%	Weekly	46.9%	ChatGPT Free Subscription	72.4%
2025	9.0%	Less than Monthly	15.5%	ChatGPT Plus	12.8%
		Monthly	12.4%	ChatGPT Pro	2.1%

Table 3: ChatGPT usage among our respondents, who mostly use Free-tier accounts and mostly use ChatGPT at least weekly.

Privacy Attitudes				
Statement	**Mean**	**SD**	**t**	**p <**
Online companies collect too much personal info	74.9	22.4	19.3	0.01
In general, I trust websites	43.8	24.6	-4.4	0.01
In general, I believe privacy is important	88.7	15.1	44.4	0.01
Privacy concerns about conversations with ChatGPT	63.6	27.9	8.5	0.01
My chats with bots like ChatGPT will stay private	40.0	27.6	-6.9	0.01

Table 4: Participants disagreed with statements that they trust websites and believe their chat data will stay private, while agreeing with statements that they have privacy concerns about ChatGPT data and that companies collect too much personal information.

Gilbert, Vitak, and Shilton (2021), we hypothesized that we would observe statistically significant agreement with questions like online companies collect too much personal information" and in general, I believe privacy is important", with means between 50 and 60. However, we observe notably higher means and tt t-values than expected.

In Table 5, we observe agreement with all three questions probing participants' concerns about AI specifically, reflecting concerns about receiving inaccurate information from AI, not understanding what AI can do, and personal data being misused by AI. These responses are high in magnitude and consistent across political demographic groups (*e.g.,* Republicans, Democrats), consistent with large-scale surveys about the extent to which AI concerns are shared across society (Faverio and Tyson 2023). Taken together, our participants' responses add to studies indicating that Americans' concern about data privacy has grown significantly over recent years (Faverio 2023), including with respect to data exchanged with AI.

Private Data Exchange

As described in Table 6, the only form of data participants were willing to share with chatbots in order to receive more personalized responses and insights was their chatbot history; participants were not willing to share access to their search history, email, or on-device data and applications in exchange for these features. These results were consistent when participants were instead presented with the option to exchange their data for a premium ($200/month) chatbot subscription, as described in Table 7. This is noteworthy given the incorporation of LLM-based chatbots into operating systems (as in the case of assistants like Microsoft's Copilot), mobile devices (as with ChatGPT for iPhone), and web interfaces, most notably AI agents capable of carrying out actions on behalf of the user. While the public could conceivably be willing to provide access to data for emerging LLM-based technologies in exchange for notable new capabilities, our results suggest that Americans are *not* willing to make that trade simply to equip a chatbot assistant with personalized features, even if those features are valued at $200.

AI Concerns Question Battery				
Concern	**Mean**	**SD**	**t**	**p <**
People getting inaccurate information from AI	72.8	21.8	18.1	0.01
People not understanding what AI can do	68.2	24.5	12.9	0.01
People's personal info being misused by AI	72.7	24.3	16.2	0.01

Table 5: Responses to "When it comes to artificial intelligence, how concerned are you about...[**Concern**]?" demonstrates consistent concerns about AI data privacy and quality.

Willingness to Share Personal Data for Personalization				
Information Type	**Mean**	**SD**	**t**	**p<**
Chatbot History	62.9	27.4	8.0	0.01
Search History	34.9	29.6	-8.4	0.01
Email	16.6	23.8	-18.7	0.01
Device	21.4	26.3	-16.5	0.01

Table 6: Responses to "I would be willing to give chatbots access to my **[Information Type]** in exchange for more personalized responses and insights" show that most respondents would not share personal data to improve a chatbot.

As shown in Table 8, respondents were ambivalent about their chatbot history being used to improve search results or product recommendations, as reflected in means just below 50 and large standard deviations. However, respondents were *not* willing to use their chatbot history to improve their social media experience. Providers like Meta and X have implemented LLM-based chatbots on their platforms (xAI 2025; Meta 2024), but our results suggest that respondents prefer to preserve the independence of their chat data from their social networks, even if it is useful for their experience

Willingness to Share Personal Data for Premium Chatbot				
Information Type	**Mean**	**SD**	**t**	**p <**
Chatbot History	61.1	33.6	5.7	0.01
Search History	36.8	33.7	-6.8	0.01
Email	24.7	32.1	-13.7	0.01
Device	26.7	32.4	-12.4	0.01

Table 7: Responses to "In exchange for a premium chatbot subscription — which is valued at $200 a month and includes priority access to new features — I would be willing to give access to my...[**Information Type**]?" show that even access to premium features does not induce respondents to share their personal data with a chatbot.

Cross-Application Integration of Chat History				
Use of Chat History	**Mean**	**SD**	**t**	**p <**
Search engine results	47.7	32.0	-1.2	0.01
Product recommendations	46.9	32.9	-1.6	0.01
Popularity of my posts on social media	26.6	30.3	-13.4	0.01
Content moderation on my social media	32.6	30.7	-9.8	0.01

Table 8: Responses to "I would be willing to have my conversation history used to improve...[**Use of Chat History**]" demonstrate little enthusiasm for integrating chat data to improve other applications, particularly on social media.

Factorial Vignettes

To analyze the 9000 factorial vignette judgments made by our survey participants, we fit two Linear Mixed Models to the 1) Appropriateness and 2) Level of Concern dependent variables. As shown in Table 9, our findings indicate that only the Consent and Anonymity factors have a significant effect on the Appropriateness variable, while the Consent and Privacy factors have a significant effect on the Level of Concern variable. Though Anonymity has a large t-value in the Concern model, its p-value narrowly misses the threshold for significance (p=.06). While results for the Privacy, Anonymity, and Consent factors align with our hypotheses, we found that our expectations about the remaining factors were not borne out. We expected that our American survey respondents would report lower levels of concern and higher perceived appropriateness for chat data exchanged with U.S. entities (and vice-versa for European Union and China), but our model indicates that there is no significant difference between the levels of the Location factor for either dependent variable. Similarly, though we expected to observe higher levels of concern when chat data was exchanged with entities such as an insurance company or a big tech company, the model indicates no statistically discernible difference between the levels of the Role factor. In keeping with our Privacy Attitudes results, the grand means of our dependent variables across all vignettes reflect a sense of concern about the exchange of chat data, as the mean for Appropriateness was 37.6 (σ=31.3), and for Concern was 67.3 (σ=30.7).

Interaction Effects. In accordance with our hypotheses, we tested interactions for responses to our Privacy Attitudes questions with 1) the Consent factor and 2) the Location factor. Note that while we fit the interactions to the full model (with all factors), we only examined results for Location and Consent in accordance with our hypotheses, and to reduce the risk of Type 1 errors. Though we hypothesized that responses to the Privacy Attitudes questions indicating greater privacy concern would interact with the Consent/Awareness factor, such that we would observe increases in the level of concern when informed consent was not obtained, we did not observe any such interaction. As clearly indicated from our main effects, Consent/Awareness exerts a significant influence on perceptions of both appropriateness and concern, but the interaction model results indicate that this is not determined by participants' self-reported privacy attitudes. We include full results for interaction effects in the Appendix.

ChatGPT Usage

Table 10 describes the most common topics discussed with ChatGPT by our respondents. More than half use ChatGPT to discuss academic or scientific topics, and to discuss topics relevant to supporting their job. More important for this study, 47.0% of respondents used ChatGPT for Health and Wellness Advice, while 35.3% used the model for Personal Finances, and 26.7% used it for Mental Health Support, and 12.7% for Legal Guidance. While information about any of the topics we asked about could conceivably be used to make valuable inferences about a user, discussions of these topics present clear vectors for the exchange of sensitive personal information. It is thus notable that so many respondents willingly discuss such sensitive topics with ChatGPT, despite the clear reservations about chatbot privacy reported in our survey's previous question batteries (*e.g.*, Privacy Attitudes).

Table 11 reports the most common tasks for which ChatGPT was used by our respondents. Searching for Information was the most common task, selected by very nearly the entirety of our sample, and perhaps indicating why, in our subsequent analysis, the perceived sensitivity of chat conversations is analogous to that of one's browsing history. We also note that, compared to other studies of user chatbot data, our participants self-report substantially lower uses of chatbots for creative work, sexual roleplay, and software development, which may speak to a gap between stated and revealed preferences (Tamkin et al. 2024; Longpre et al. 2024).

Chat Data Sensitivity

Our respondents' perceived sensitivity of various forms of personal information are described in Figure 2. A Friedman test revealed a statistically significant difference in perceived sensitivity between information types, with $\chi^2(13) = 2160.5$, $p < 0.01$. Post-hoc tests with Bonferroni corrections further indicated significant differences between the perceived sensitivity of chatbot conversations and each of the other forms of personal information, with the notable exception of records of websites visited online (*i.e.,* browsing history), for which the post-hoc comparison was not significant. Respondents were most likely to rank their Social Security number, details of their physical location over a period

Factors	Appropriateness LMM				Concern LMM			
	Est	*SE*	*t*	*p*	Est	*SE*	*t*	*p*
Intercept	41.38	1.69	24.48	<0.01	62.36	1.76	35.40	<0.01
Role (Intercept: Hospital)	Est	*SE*	*t*	*p*	Est	*SE*	*t*	*p*
Big tech company	1.28	1.07	1.20	0.23	0.18	1.12	0.16	0.87
University computer scientist	1.04	1.15	0.90	0.37	0.92	1.08	0.85	0.39
University social science researcher	0.31	1.05	0.29	0.77	1.54	1.15	1.34	0.18
Charitable Foundation	0.12	1.16	0.11	0.92	1.61	1.21	1.32	0.19
Government agency	-0.23	1.09	-0.21	0.83	0.69	1.16	0.59	0.55
Insurance company	0.32	1.12	0.29	0.77	1.26	1.10	1.15	0.25
Purpose (Intercept: Improving user experience with AI)	Est	*SE*	*t*	*p*	Est	*SE*	*t*	*p*
Assessing mental health	-0.95	1.07	-0.87	0.38	1.02	1.07	0.95	0.34
Creating a public dataset for AI research	0.13	1.05	0.12	0.90	0.39	1.09	0.35	0.72
Fighting Terrorism	-1.41	1.11	-1.27	0.21	0.50	1.12	0.45	0.66
Personalizing Advertising	-0.71	1.17	-0.61	0.54	0.08	1.17	0.07	0.94
Predicting human behavior	-0.47	1.13	-0.42	0.68	1.04	1.16	0.89	0.37
Studying AI risks to society	0.41	1.09	0.37	0.71	0.02	1.08	0.02	0.98
Training future AI models	-0.38	1.16	-0.33	0.74	0.54	1.13	0.48	0.63
Content Type (Intercept: All conversations with ChatGPT)	Est	*SE*	*t*	*p*	Est	*SE*	*t*	*p*
Conversations using ChatGPT to help with your job	0.03	0.94	0.03	0.97	0.88	0.89	0.98	0.33
Conversations about social life and personal relationships	-1.50	0.82	-1.82	0.07	1.31	0.88	1.49	0.14
Conversations about your personal health and wellness	-0.44	0.86	-0.51	0.61	1.26	0.89	1.42	0.16
Conversations where you sought legal or ethical guidance	1.47	1.01	1.46	0.14	0.17	1.00	0.17	0.87
Location (Intercept: United States)	Est	*SE*	*t*	*p*	Est	*SE*	*t*	*p*
China	1.28	1.07	1.20	0.23	0.81	0.70	1.15	0.25
The European Union	1.04	1.15	0.90	0.37	-0.02	0.73	-0.02	0.98
Consent (Intercept: Asked for consent in advance)	Est	*SE*	*t*	*p*	Est	*SE*	*t*	*p*
Not informed that your data was collected	-4.15	0.76	-5.46	**<.001**	2.42	0.73	3.34	**<.001**
Informed that your data was collected	-0.37	0.71	-5.42	0.60	0.23	0.71	0.32	0.75
Anonymity (Intercept: Anonymized)	Est	*SE*	*t*	*p*	Est	*SE*	*t*	*p*
Not anonymized	-1.36	0.58	-2.36	**<.001**	1.12	0.58	1.92	0.06
Privacy (Intercept: PII Removed)	Est	*SE*	*t*	*p*	Est	*SE*	*t*	*p*
PII not removed	-2.03	0.59	-3.48	**<.001**	2.33	0.58	3.98	**<.001**

Table 9: Significant LMM results for Consent, Anonymity, and Privacy with Appropriateness, and Consent and Privacy with Concern, indicate that variation in perceptions of chatbot information flows depends primarily on the Transmission Principle.

of time, the state of their health, and the content of their texts and phone conversations as sensitive or highly sensitive. Though chatbot conversations were less likely to be rated as highly sensitive than these forms of information, more than 80% of respondents rated chatbot conversations as sensitive or highly sensitive, with the majority (60%) characterizing this data as sensitive. Chatbot conversations were perceived as more sensitive than the contents of email messages or social media posts, but less sensitive than text and phone conversations, suggesting chatbot interactions may occupy an intermediate space between information exchanges that are more formal (email) or performative (social media) and those that are more private and personal (text and phone).

Discussion

Our research reflects both growing concern about data privacy online and declining trust for data protections (Faverio and Tyson 2023; Faverio 2023). The societal consequences are identifiable in our findings. Consider that *none* of the levels for our Role factor exhibited significant effects on appropriateness or concern, such that we observe no benevolence paid to ostensibly responsible actors such as university researchers or charitable foundations. Similarly, we observed no significant effects for the Purpose factor, such that respondents were not swayed when scenarios included studying AI risks, fighting terrorism, or creating public datasets. We also found no support for our hypothesis that American respondents would be more concerned about foreign (China and European) receivers of their chat data than domestic ones. Our results suggest participants view chatbot

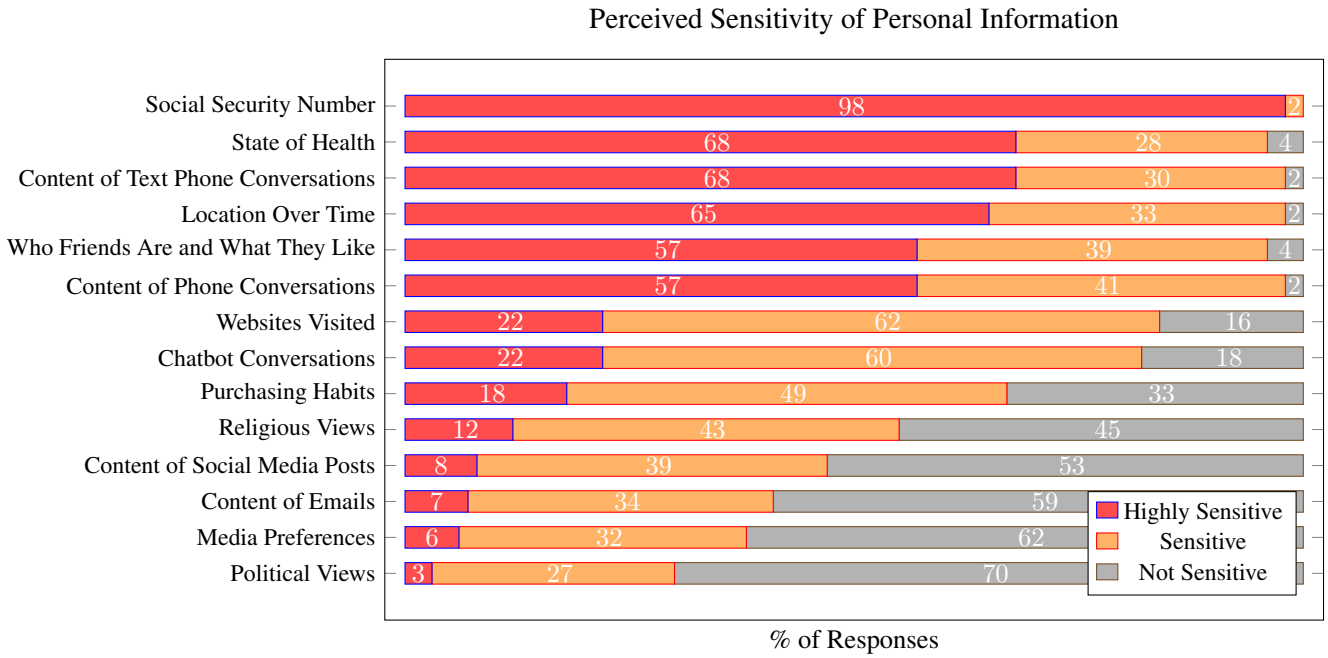

Figure 2: 82% of respondents viewed chatbot data as sensitive or highly sensitive. Chat data was perceived as more sensitive than email contents, social media posts, and religious or political views, but less sensitive than text or verbal phone conversations.

Topics Discussed with ChatGPT		
Topic	**Count**	**Percentage**
Work/Job Support	159	53 %
Scientific or Academic Information	159	53 %
Health or Wellness Advice	141	47 %
Entertainment	119	40 %
Current Events	118	40 %
Hobbies	106	35 %
Travel	106	35 %
Personal Finances	106	35 %
Mental Health Support	80	26 %
Personal Relationships	62	21 %
Politics	57	19 %
Legal Guidance	38	13 %
Religion	28	9 %
Sexual & Erotic	17	6 %
Other	20	7 %

Table 10: Substantial pluralities of participants discussed sensitive topics such as health and wellness, personal finances, mental health, and legal guidance with ChatGPT, despite their reservations about the privacy of this data.

Tasks ChatGPT is Used For		
Task	**Count**	**Percentage**
Searching for Information	288	96 %
Writing and Editing	196	65 %
Asking for Advice	155	52 %
Personal Education (learning)	143	48 %
Daily Life Skills (cooking, finances)	89	30 %
Computer Programming	87	30 %
Art and Design	49	16 %
Other	20	7 %

Table 11: 96% of respondents used ChatGPT to search for information, suggesting a reason that respondents also rate chat data as having similar sensitivity to browsing history.

data sharing as problematic, regardless of content, recipient, or stated purpose, particularly when exchanges occur without informed consent, anonymization, and proper removal of sensitive data. We acknowledge that meaningful distinctions for users may exist that differ from those captured in our experimental design, which focuses on hypothetical vignettes. But on average, our participants *disagreed* with "My chats with bots like ChatGPT will stay private," and *agreed* they were concerned about personal information being misused by AI. Together, our findings reflect consistent distrust in the management and collection of personal information, including with LLM-based chatbots. Our survey data indicate users are unwilling to provide chatbots access to emails, search history, or devices in exchange for better services—including premium-tier ($200/month) subscriptions. Similarly, participants were unwilling to have chatbot data used to improve their online experience, especially on social media. These results depict general skepticism about chatbot data integration relevant to AI companies and regulators. Agentic AI systems promise to execute complex actions on behalf of users (Acharya, Kuppan, and Divya 2025; Kapoor et al. 2024), and while early demonstrations are technically impressive, they depend upon extensive sensitive user data

Research Question	Finding	Implication
RQ1: How do users perceive the sensitivity of chatbot data relative to other forms of personal information?	80+% of respondents view chat data as sensitive (less sensitive than a phone call but more sensitive than an email).	Chat data is disclosive, and a gap in reported sensitivity may indicate users' inability to recognize privacy tradeoffs.
RQ2: What factors influence users' concern with handling of their chat data?	Consent, anonymization, and removal of PII (*i.e.,* features of a transmission principle in contextual integrity) are significant in our model.	For an emerging technology like LLM-based chatbots, users judge information exchanges based on governing principles that are well understood and often used.
RQ3: Are users willing to exchange their personal data for premium or personalized chatbot features?	No, respondents reject exchanging data (emails, search histories, device access) for premium and personalized AI features.	AI companies should not assume willingness to integrate data across contexts for more advanced models (*e.g.*, agentic AI).

Table 12: A Summary of Our Primary Research Questions, Findings, and Implications.

that users seem unwilling to trade. We establish a straightforward starting point for investigating consumer tradeoffs between privacy and advanced AI by showing $200.00 in value is insufficient for unrestricted agentic data access.

Such unwillingness makes sense if adherence to privacy norms is both scarce and valuable in an online economy driven by user data (Lammi and Pantzar 2019). In contextual integrity, the three significant factors isolated using our LMMs (informed consent, anonymity, and privacy) align with "transmission principles," or aspects of information exchange governed by expectations and rights of data subjects (Malkin 2022). This suggests that for emerging technology like LLM-based chatbots, users judge appropriateness based on well-understood governing principles. Our participants signaled the most meaningful aspect when considering responsible data use was simply being asked in advance (consent) and having personal information removed.

Our focus on transmission principles may reflect a post-hoc truth (*i.e. that privacy and consent matter)*, but our results also point to emerging areas of interest to ethics researchers and policymakers (Susser 2019). Respondents described chatbot data as sensitive and expressed doubt it would remain private, and large pluralities report regularly discussing sensitive topics like personal finances, legal guidance, and health with chatbots. The ethical use of emerging technologies is more complicated than pointing to a privacy paradox (Martin and Nissenbaum 2016; Dielin and Trepte 2015), but growing work demonstrates users are unaware of how much sensitive data they exchange with chatbots, how this might impact their lives if disclosed, and ramifications for having chat data seized or used adversarially by law enforcement. Moreover, while respondents value informed consent, inconsistencies between privacy concerns and disclosure behaviors, combined with chatbots' complexity and opacity, suggest meaningful consent is challenging to achieve (Winograd 2022; Atata 2024). Consequently, traditional notice-and-consent paradigms, shown ineffective in protecting online privacy (Barocas and Nissenbaum 2009; Hijjawi 2024), may be inadequate for addressing chatbot privacy concerns. We believe our results establish a significant baseline about factors that matter to end-user privacy notions and value users assign to chatbot interactions. We summarize findings in Table 12.

Limitations

One of the central limitations of our approach is that we restricted our participant population to people residing in the U.S. This was primarily due to financial constraints, and we are in the process of expanding this work to a comparative study of AI and chatbot privacy norms around the world. Moreover, as described in the Appendix, we deviated from our preregistration, most notably by adding our questions about personalization, privacy valuation, and integration of chat data. We thus note that results from these new questions should *not* be viewed as having passed the "severe" test of preregistration. However, because we were also *more* conservative with our data analysis than declared, future work might examine non-significant but marginal effects identified by our study, especially for Content and Location, as we believe we are more likely to make Type 2 errors, rejecting true hypotheses. Finally, future work might further develop our findings within the contextual integrity framework. Though we situate our results within the theory of contextual integrity, our methods cannot isolate more complex information flows without increasing the chances of making Type 1 errors.

Conclusion

We set out to understand the emerging privacy norms around chatbots. Our work demonstrated both broad concern about the use of chatbot data, particularly when data is integrated across contexts, as well as the primacy of informed consent, anonymity, and privacy (associated with transmission principle in contextual integrity) for determining the appropriateness of information flows with LLM-based chatbots.

Data Availability Statement

The data and software developed for this study have been archived and are openly available in Dataverse: https://doi.org/10.7910/DVN/M6ABJ3 and Github https://github.com/WeberLab-UW/chatbot-privacy

Appendices

The appendix to this article is available in the arXiv version of this paper: https://doi.org/10.48550/arXiv.2508.06760

Positionality Statement

This research highlights public concerns surrounding the privacy risks attendant to a novel emerging technology, LLM-based chatbots. Given the multidisciplinary nature of our work, we sought to include the expertise of scholars from many disciplines on this research, including those who characterize their primary research domains as social science, computer science, statistics, and public policy. In addition to their intellectual diversity, the authors of this work are also demographically diverse, reflective of a wide array of racial, ethnic, gender, and national identities. We believe that these forms of diversity ultimately benefited the paper.

Ethical Impact

We do not anticipate immediate harms from the findings of this paper, as we focus our attention on the privacy concerns of Americans, and hope that our work will be used to foster more responsible design and public policy with respect to LLM-based chatbots. By focusing on Americans, we acknowledge that we exclude many important perspectives; however, as noted in the Limitations section, we expect to soon expand this work to several locations around the world.

Ethical Statement

Theories of contextual integrity have long sought to bridge the divide between social science and ethical theory as it pertains to the modern information economy. While the results of this work are primarily empirical and belong to the domain of social science, our research contributes perspectives that can inform normative ethical work with respect to the responsible handling of data produced in interaction with AI chatbots, assistants, and agents.

Acknowledgements

This research was supported by grants from the Sloan Foundation (G-2018-11217) and the Institute for Museum and Library Services (RE-252290-OLS-22).

References

Acharya, D. B.; Kuppan, K.; and Divya, B. 2025. Agentic AI: Autonomous Intelligence for Complex Goals–A Comprehensive Survey. *IEEE Access*.

Act, E. A. I. 2024. The EU Artificial Intelligence Act.

Ali, M.; Arunasalam, A.; and Farrukh, H. 2025. Understanding Users' Security and Privacy Concerns and Attitudes Towards Conversational AI Platforms. *arXiv preprint arXiv:2504.06552*.

Alkamli, S.; and Alabduljabbar, R. 2024. Understanding privacy concerns in ChatGPT: A data-driven approach with LDA topic modeling. *Heliyon*, 10(20).

Apthorpe, N.; Shvartzshnaider, Y.; Mathur, A.; Reisman, D.; and Feamster, N. 2018. Discovering smart home internet of things privacy norms using contextual integrity. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 2(2): 1–23.

Arntz, P. 2024. AI chatbot provider exposes 346,000 customer files, including ID documents, resumes, and medical records. https://www.malwarebytes.com/blog/news/2024/12/ai-chatbot-provider-exposes-346000-customer-files-including-id-documents-resumes-and-medical-records. [Accessed 05-17-2025].

Atata, B. B. 2024. The AI Privacy Paradox: A Comparative Analysis of EU and US Approaches to Regulating Artificial Intelligence and Protecting Personal Data. *International Research Journal of Modernization in Engineering Technology and Science*, 6: 1–15.

Barocas, S.; and Nissenbaum, H. F. 2009. On Notice: The Trouble with Notice and Consent. In *Proceedings of the Engaging Data Forum: The First International Forum on the Application and Management of Personal Electronic Information*.

Barth, A.; Datta, A.; Mitchell, J. C.; and Nissenbaum, H. 2006. Privacy and contextual integrity: Framework and applications. In *2006 IEEE symposium on security and privacy (S&P'06)*, 15–pp. IEEE.

Beaty, A. 2025. ChatGPT's user base just doubled in 6 months - to more than 400 million weekly users. https://www.zdnet.com/article/chatgpts-user-base-just-doubled-in-6-months-to-more-than-400-million-weekly-users/. [Accessed 05-23-2025].

Belen Saglam, R.; Nurse, J. R.; and Hodges, D. 2021. Privacy concerns in chatbot interactions: When to trust and when to worry. In *HCI International 2021-Posters: 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part II 23*, 391–399. Springer.

Burgess, M.; and Newman, L. 2025. DeepSeek's Popular AI App Is Explicitly Sending US Data to China. https://www.wired.com/story/deepseek-ai-china-privacy-data/. [Accessed 05-23-2025].

Cameron, A. C.; Miller, D. L.; et al. 2010. Robust inference with clustered data. *Handbook of empirical economics and finance*, 106: 1–28.

Chan, A.; Ezell, C.; Kaufmann, M.; Wei, K.; Hammond, L.; Bradley, H.; Bluemke, E.; Rajkumar, N.; Krueger, D.; Kolt, N.; et al. 2024. Visibility into AI agents. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 958–973.

Chan, A.; Salganik, R.; Markelius, A.; Pang, C.; Rajkumar, N.; Krasheninnikov, D.; Langosco, L.; He, Z.; Duan, Y.; Carroll, M.; et al. 2023. Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 651–666.

Chen, J.; Liu, Z.; Huang, X.; Wu, C.; Liu, Q.; Jiang, G.; Pu, Y.; Lei, Y.; Chen, X.; Wang, X.; et al. 2024. When large language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web*, 27(4): 42.

Cheng, Z.; Wan, D.; Abueg, M.; Ghalebikesabi, S.; Yi, R.; Bagdasarian, E.; Balle, B.; Mellem, S.; and O'Banion, S. 2024. CI-Bench: Benchmarking Contextual Integrity

of AI Assistants on Synthetic Data. *arXiv preprint arXiv:2409.13903*.

Cox, J. 2025. Nearly 100,000 chatgpt conversations were searchable on google.

Dielin, T.; and Trepte, S. 2015. Is the privacy paradox a relic of the past? An in-depth analysis of privacy attitudes and privacy behaviors. *European Journal of Social Psychology*, 45(3): 285–297.

Fan, W.; Li, H.; Deng, Z.; Wang, W.; and Song, Y. 2024. Goldcoin: Grounding large language models in privacy laws via contextual integrity theory. *arXiv preprint arXiv:2406.11149*.

Faverio, M. 2023. Key findings about Americans and data privacy. https://www.pewresearch.org/short-reads/2023/10/18/key-findings-about-americans-and-data-privacy/. [Accessed 05-17-2025].

Faverio, M.; and Tyson, A. 2023. What the data says about Americans' views of artificial intelligence. https://www.pewresearch.org/short-reads/2023/11/21/what-the-data-says-about-americans-views-of-artificial-intelligence/. [Accessed 05-17-2025].

Ghalebikesabi, S.; Bagdasaryan, E.; Yi, R.; Yona, I.; Shumailov, I.; Pappu, A.; Shi, C.; Weidinger, L.; Stanforth, R.; Berrada, L.; et al. 2024. Operationalizing contextual integrity in privacy-conscious assistants. *arXiv preprint arXiv:2408.02373*.

Gilbert, S.; Vitak, J.; and Shilton, K. 2021. Measuring Americans' comfort with research uses of their social media data. *Social Media+ Society*, 7(3): 20563051211033824.

Golda, A.; Mekonen, K.; Pandey, A.; Singh, A.; Hassija, V.; Chamola, V.; and Sikdar, B. 2024. Privacy and security concerns in generative AI: a comprehensive survey. *IEEE Access*.

Hartman, R.; Moss, A. J.; Jaffe, S. N.; Rosenzweig, C.; Litman, L.; and Robinson, J. 2023. Introducing Connect by CloudResearch: Advancing online participant recruitment in the digital age.

Hijjawi, F. 2024. *The End of Consent: Data and the Corporate-Consumer Relationship*. Master's thesis.

Kapoor, S.; Stroebl, B.; Siegel, Z. S.; Nadgir, N.; and Narayanan, A. 2024. Ai agents that matter. *arXiv preprint arXiv:2407.01502*.

Kumar, P. C.; Zimmer, M.; and Vitak, J. 2024. A roadmap for applying the contextual integrity framework in qualitative privacy research. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1): 1–29.

Lakens, D. 2024. When and how to deviate from a preregistration. *Collabra: Psychology*, 10(1).

Lammi, M.; and Pantzar, M. 2019. The data economy: How technological change has altered the role of the citizen-consumer. *Technology in Society*, 59: 101157.

Li, H.; Fan, W.; Chen, Y.; Cheng, J.; Chu, T.; Zhou, X.; Hu, P.; and Song, Y. 2024. Privacy checklist: Privacy violation detection grounding on contextual integrity theory. *arXiv preprint arXiv:2408.10053*.

Longpre, S.; Mahari, R.; Lee, A.; Lund, C.; Oderinwale, H.; Brannon, W.; Saxena, N.; Obeng-Marnu, N.; South, T.; Hunter, C.; et al. 2024. Consent in crisis: The rapid decline of the ai data commons. *Advances in Neural Information Processing Systems*, 37: 108042–108087.

Ma, R.; Maidhof, C.; Carrillo, J. C.; Lindqvist, J.; and Such, J. 2025. Privacy Perceptions of Custom GPTs by Users and Creators. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–18.

Malkin, N. 2022. Contextual integrity, explained: A more usable privacy definition. *IEEE Security & Privacy*, 21(1): 58–65.

Martin, K.; and Nissenbaum, H. 2016. Measuring privacy: An empirical test using context to expose confounding variables. *Colum. Sci. & Tech. L. Rev.*, 18: 176.

Martin, K. E. 2012. Diminished or just different? A factorial vignette study of privacy as a social contract. *Journal of Business Ethics*, 111: 519–539.

Matz, S. C.; Teeny, J. D.; Vaid, S. S.; Peters, H.; Harari, G. M.; and Cerf, M. 2024. The potential of generative AI for personalized persuasion at scale. *Scientific Reports*, 14(1): 4692.

Meta. 2024. Meet Your New Assistant: Meta AI, Built With Llama 3. https://about.fb.com/news/2024/04/meta-ai-assistant-built-with-llama-3/. [Accessed 04-28-2024].

Mireshghallah, N.; Antoniak, M.; More, Y.; Choi, Y.; and Farnadi, G. 2024. Trust no bot: Discovering personal disclosures in human-llm conversations in the wild. *arXiv preprint arXiv:2407.11438*.

Mireshghallah, N.; Kim, H.; Zhou, X.; Tsvetkov, Y.; Sap, M.; Shokri, R.; and Choi, Y. 2023. Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity Theory. *ArXiv*, abs/2310.17884.

Mok, C. 2025. Taking Stock of the DeepSeek Shock. https://cyber.fsi.stanford.edu/publication/taking-stock-deepseek-shock. [Accessed 05-23-2025].

Mussgnug, A. M. 2024. Technology as uncharted territory: Contextual integrity and the notion of AI as new ethical ground. *arXiv preprint arXiv:2412.05130*.

Nasr, M.; Carlini, N.; Hayase, J.; Jagielski, M.; Cooper, A. F.; Ippolito, D.; Choquette-Choo, C. A.; Wallace, E.; Tramèr, F.; and Lee, K. 2023. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*.

Nasr, M.; Rando, J.; Carlini, N.; Hayase, J.; Jagielski, M.; Cooper, A. F.; Ippolito, D.; Choquette-Choo, C. A.; Tramèr, F.; and Lee, K. 2025. Scalable extraction of training data from aligned, production language models. In *The Thirteenth International Conference on Learning Representations*.

Ngong, I.; Kadhe, S.; Wang, H.; Murugesan, K.; Weisz, J. D.; Dhurandhar, A.; and Ramamurthy, K. N. 2025. Protecting users from themselves: Safeguarding contextual privacy in interactions with conversational agents. *arXiv preprint arXiv:2502.18509*.

Nissenbaum, H. 2004. Privacy as contextual integrity. *Wash. L. Rev.*, 79: 119.

Nissenbaum, H. 2019. Contextual integrity up and down the data food chain. *Theoretical inquiries in law*, 20(1): 221–256.

Nosek, B. A.; Ebersole, C. R.; DeHaven, A. C.; and Mellor, D. T. 2018. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11): 2600–2606.

OpenAI. 2024. Introducing ChatGPT Pro. *OpenAI Blog*, (): .

Palmer, A.; Smith, N. A.; and Spirling, A. 2024. Using proprietary language models in academic research requires explicit justification. *Nature Computational Science*, 4(1): 2–3.

Paris, T.; Moon, A.; and Guo, J. 2025. Opening the Scope of Openness in AI. *arXiv preprint arXiv:2505.06464*.

Riedemann, L.; Labonne, M.; and Gilbert, S. 2024. The path forward for large language models in medicine is open. *npj Digital Medicine*, 7(1): 339.

RyokoAI. 2023. Dataset Card for ShareGPT90K. https://huggingface.co/datasets/RyokoAI/ShareGPT52K. [Accessed 05-17-2025].

Solaiman, I. 2023. The gradient of generative AI release: Methods and considerations. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, 111–122.

Susser, D. 2019. Notice after notice-and-consent: Why privacy disclosures are valuable even if consent frameworks aren't. *Journal of Information Policy*, 9: 148–173.

Tamkin, A.; McCain, M.; Handa, K.; Durmus, E.; Lovitt, L.; Rathi, A.; Huang, S.; Mountfield, A.; Hong, J.; Ritchie, S.; et al. 2024. Clio: Privacy-Preserving Insights into Real-World AI Use. *arXiv preprint arXiv:2412.13678*.

Vitak, J.; and Zimmer, M. 2023. Surveillance and the future of work: exploring employees' attitudes toward monitoring in a post-COVID workplace. *Journal of Computer-Mediated Communication*, 28(4): zmad007.

Winograd, A. 2022. Loose-lipped large language models spill your secrets: The privacy implications of large language models. *Harv. JL & Tech.*, 36: 615.

Wolfe, R.; and Mitra, T. 2024a. The impact and opportunities of Generative AI in fact-checking. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1531–1543.

Wolfe, R.; and Mitra, T. 2024b. The Implications of Open Generative Models in Human-Centered Data Science Work: A Case Study with Fact-Checking Organizations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 1595–1607.

xAI. 2025. Grok 3 Beta — The Age of Reasoning Agents. https://x.ai/news/grok-3. [Accessed 05-23-2025].

Zhang, Z.; Jia, M.; Lee, H.-P.; Yao, B.; Das, S.; Lerner, A.; Wang, D.; and Li, T. 2024. "It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–26.

Zhao, W.; Ren, X.; Hessel, J.; Cardie, C.; Choi, Y.; and Deng, Y. 2024. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Li, T.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Li, Z.; Lin, Z.; Xing, E. P.; et al. 2023. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *arXiv preprint arXiv:2309.11998*.