PROTECTING USERS FROM THEMSELVES: SAFEGUARDING CONTEXTUAL PRIVACY IN INTERACTIONS WITH CONVERSATIONAL AGENTS

Ivoline Ngong[*], **Swanand Kadhe, Hao Wang, Keerthiram Murugesan, Justin D. Weisz, Amit Dhurandhar, Karthikeyan Natesan Ramamurthy**

IBM Research.

kngongiv@uvm.edu,

{swanand.kadhe,hao,keerthiram.murugesan}@ibm.com,

{jweisz,adhuran,knatesa}@us.ibm.com

ABSTRACT

Conversational agents are increasingly woven into individuals' personal lives, yet users often underestimate the privacy risks involved. The moment users share information with these agents (e.g., LLMs), their private information becomes vulnerable to exposure. In this paper, we formalize the notion of *contextual privacy* for user interactions with LLMs. It aims to minimize privacy risks by ensuring that users (sender) disclose only information that is both relevant and necessary for achieving their intended goals when interacting with LLMs (untrusted receiver). Through a formative design user study, we observe how even "privacy-conscious" users inadvertently reveal sensitive information through indirect disclosures. Based on insights from this study, we propose a system that operates between users and LLMs. The system identifies potentially sensitive information in user prompts and suggests reformulations to ensure that only contextually relevant information is shared. Using examples from ShareGPT, we demonstrate how users often disclose private information beyond their intent and illustrate how our system can guide them toward more privacy-preserving interactions with LLMs while still achieving their intended outcomes.

1 INTRODUCTION

Conversational agents such as large language model (LLM)-based chatbots are a double-edged sword. In specialized systems such as customer service platforms and medical assistants, they offer valuable services to individual users (Mariani et al., 2023; Kumar et al., 2024b; Yang et al., 2023; Chow et al., 2023; Rani et al., 2024; Sadhu et al., 2024). However, these models present unique privacy challenges that fundamentally differ from human-human interactions. For example, they can memorize (Carlini et al., 2019; Biderman et al., 2024; McCoy et al., 2023; Zhang et al., 2023) and potentially misuse information (Kumar et al., 2024a). They are vulnerable to data breaches or unauthorized sharing with third parties (Nagireddy et al., 2024; Carlini et al., 2021; Nasr et al., 2023), and user-provided data may be incorporated into future model training, potentially resulting in unintended information leaks during deployment (Zanella-Béguelin et al., 2020). While existing research explores techniques like unlearning to address post-training privacy concerns, we focus on a critical but understudied aspect: helping users make informed decisions about what information they share with these untrusted agents in the first place. This is particularly important because once information is shared with an LLM, users lose control over how it might be used or disseminated.

Many users are unaware of the privacy risks associated with their interactions with LLMs. As these models become more adept at handling complex tasks and users remain uninformed about privacy

[*]Graduate student at University of Vermont. Work done during summer internship at IBM Research.

risks, they develop increasing trust in both the technology and their own ability to protect themselves (Natarajan & Gombolay, 2020; Zhang et al., 2024; Mireshghallah et al., 2024; Cummings et al., 2023; Dou et al., 2023). In our formative user study, we found that users often believe they can protect their private information by keeping conversations vague and removing obvious Personal Identifiable Information (PII). However, when shown examples of how indirect disclosures could reveal sensitive details in specific contexts, many participants began to question their privacy protection strategies. The participants expressed a desire for a real-time system that could highlight privacy risks and assist in revising information before sharing it with conversational agents. This motivates the main objective of this paper:

Develop a system that operates between users and conversational agents
to detect and manage sensitive and contextually inappropriate information during interactions.

In this paper, we explore how contextual integrity theory (Nissenbaum, 2004; 2011) can guide privacy protection when users interact with LLMs. Contextual integrity defines privacy not merely as hiding personal information, but as maintaining appropriate information flows within specific contexts. Interactions with LLMs require a careful privacy consideration because LLMs are untrusted receivers that can store, reuse, or leak information in ways users cannot control. We adapt the data minimization principle—commonly applied to organizations through regulations like the GDPR (Voigt & Von dem Bussche, 2017), which states that *"all collected data shall be adequate, relevant, and limited to what is necessary in relation to the purposes for which they are processed"*. We introduce *contextual privacy* for the specific case of `User`→`LLM` information flows. It requires that user prompts include only information that is both relevant and necessary to achieve the user's intended goals when interacting with LLMs. For instance, if a user seeks advice on managing personal finances, sharing the names of family members is unnecessary, whereas a general overview of their financial situation may be essential. Similarly, when asking for advice on managing mild seasonal allergy symptoms, details such as the user's full name or date of birth are unnecessary, while information about symptoms or lifestyle may be relevant.

To demonstrate the practical importance of our approach, we analyze conversations from a real-world dataset – ShareGPT (Chiang et al., 2023) – to identify conversations that violate contextual privacy norms. Our analysis reveals that users often share information beyond what their context requires, inadvertently exposing sensitive details that were unnecessary for their intended goals (see examples in Table 1). This finding underscores the importance of raising user awareness about privacy risks stemming from contextually inappropriate disclosures. Through our formative design study, we observe how even "privacy-conscious" participants unintentionally share sensitive information through indirect disclosures with conversational agents. The study helps us identify technical requirements to address contextual privacy violations in `User`→`LLM` information flows.

We design a system that can protect users during their interactions with conversational agents. By analyzing user inputs, detecting potentially sensitive irrelevant content, and guiding users to reformulate prompts based on contextual relevance, our system empowers users to make more informed, privacy-conscious decisions in real-time. Rather than enforcing rigid privacy rules, the system helps users understand the privacy implications of their choices while preserving their intended interaction goals.

We implement a simplified version of this system and demonstrate through selected ShareGPT examples how users can maintain appropriate information flows while achieving their conversational goals. This system represents a novel application of contextual integrity theory, focusing specifically on protecting user privacy at the point of information disclosure to LLMs.

Our main contributions include:

- formalizing contextual privacy for the specific case of `User`→`LLM` information flows, where users act as senders and LLMs as untrusted receivers,
- demonstrating through real-world examples how users unintentionally violate contextual privacy in interactions with LLMs,
- developing a system that helps users identify and reformulate potentially sensitive information while maintaining their intended goals.

Table 1: Examples of contextual privacy violations in the ShareGPT dataset. Non-essential information that should be protected is highlighted in red, illustrating cases where unnecessary sensitive details were disclosed during interactions.

User Intent	**User Prompt**
Looking for a job	My friend Justin, who was just laid off from Google, is looking for a job where he can use ML and Python. Do you have any advice for him?
Pros and cons of running	I plan to go running at 18:30 today with Pauline and Guillaume around île de la grande jatte in Levallois, France. Give me the most likely negative outcome and the most likely positive outcome of this event.
Cost of monthly medical checkup	Jing's son has recently been diagnosed with type 1 diabetes which, according to him, will cost him an extra $200 per month. How much extra will a monthly medical checkup cost?
Write Poem	Please write a valentine's day themed poem for my wife Chris. Include our 13 week old daughter named Magnolia and add in some humor.

While our proof-of-concept implementation focuses on demonstrating feasibility, it opens up new directions for research in user-centric privacy protection for LLM interactions.

2 Threat Models and Privacy Definition

In user interactions with conversational agents, privacy goes beyond simply protecting Personal Identifiable Information (PII) (Mireshghallah et al., 2024; Zhang et al., 2024), and involves ensuring that shared information is contextually appropriate, relevant, and necessary for the task at hand. We define privacy based on the theory of contextual integrity and outline the key threat models that arise during these interactions, providing a foundation for our framework.

Threat Models: Consider a scenario where users interact with large, remote, and untrusted agents such as LLM-based chatbots through APIs. These agents can be web-based or hosted on cloud-based services or private networks, and may be either general-purpose or domain-specific. Users often share personal, financial, or medical information without clear knowledge of how their data is managed, increasing privacy risks due to the lack of transparency around these agents.

We focus on a threat model where users unintentionally compromise their privacy by oversharing information. Our approach targets this *self-disclosure* threat model by guiding users to share only contextually necessary information. By identifying and highlighting unnecessary or sensitive disclosures in real-time, we assist users in controlling the information they reveal, thereby reducing the risk of unintentional privacy breaches.

Contextual Integrity in Conversational Agents: Contextual integrity (CI) offers a comprehensive framework for understanding privacy, focusing not merely on whether information is shared but on whether the sharing aligns with the norms of a given social or institutional context (Nissenbaum, 2004). CI evaluates whether the information flow adheres to appropriate standards based on the specific circumstances of the interaction. The key parameters of CI and their definitions and considerations when applied to interactions between users and conversational agents is presented in Table 2.

Based on the tenets of CI, we theoretically formalize contextual privacy for a user interacting with a conversational agent (details in Appendix C). The key idea is to identify *primary context* (which captures user's intent and the key task) from the user's query along with the prior conversation history, and determine two types of attributes in the query: (a) details that are necessary to answer the query, and (b) sensitive details that are not essential for answering the query. A contextually private user query does not contain any nonessential sensitive attributes.

3 Contextual Privacy Violations in Real-World User Interactions

We evaluated real-world conversations from the ShareGPT (Chiang et al., 2023) dataset for contextual privacy violations. To instantiate our formal privacy definition, we used an LLM-as-a-judge (specifi-

CI Entity	Definition	Function/Considerations
Sender (self)	The user sending information to the agent to achieve a task.	Ensure the user shares only relevant and necessary information.
Subject	The individual(s) about whom information is shared (self, others, or both).	Protect the privacy of the subject by identifying whether the subject is the user or another person. Information shared should respect the subject's privacy.
Receiver (agent)	The agent that receives and processes information.	Treat agent as untrusted. Apply strict privacy controls to prevent oversharing. May be domain-specific (e.g., MedicalChat Assistant) or general-purpose (e.g., ChatGPT).
Context(data type)	The broader domain or user intent (e.g., medical, finance, work-related) guiding the interaction.	Guides what information is relevant to share. In domain-specific apps, the context is predefined; in general-purpose apps, intent detection is used. Optionally, users may specify sensitive contexts.
Transmission Principle	The rule governing the flow of information between sender and receiver.	Share only essential and relevant information for the task, avoiding unnecessary or sensitive information. Respect the privacy expectations defined by context and actors.

Table 2: Entities associated with contextual integrity in conversational agents.

cally, `Meta-Llama-3.1-405B-Instruct`), and designed a prompt for the model, guided by the privacy definition (Appendix C). To manage inference costs, we selected 2,400 random single-turn conversations for analysis. For the user prompt in each conversation, the judge model determined the primary context, identified sensitive information, determined sensitive details non-essential for the task completion, and based on these, classified every prompt as either private or revealing. Our evaluation demonstrated that 18.3% of these conversations (440 out of 2400) exhibited contextually revealing privacy violations. See examples in Table 1. Manual inspection of the results generated by the judge model for consistency and correctness showed that the judge model performs a good job at classification with fairly small number of false positives and false negatives.

4 A Framework for Safeguarding Contextual Integrity

User Study to Guide Safeguarding Framework Design: We conducted a *Wizard-of-Oz* formative user study to explore how participants manage privacy when interacting with conversational agents and to identify technical requirements for our system. Following established practices in early-stage interface design research (Nielsen, 2000; Budiu, 2021; Nielsen & Landauer, 1993) where 5 participants are typically sufficient to identify major design insights, we conducted our study with six participants from our institution who were familiar with LLMs. Using three mid-fidelity UX mockups (see Appendix D.1), we probed participants on their privacy concerns, reactions to privacy disclosures, and preferences for managing sensitive information. Each mockup simulated interactions where PII and sensitive information were detected and flagged. Participants provided feedback on different approaches to identifying, flagging, and reformulating sensitive information. Insights from this formative phase shaped several key design aspects of our system, including distinguishing between necessary and unnecessary sensitive information, real-time feedback, user control over reformulations, transparency around how sensitive information is handled and flagged. The participants rated the overall approach of the system highly, with a min and max rating of 7/10 and 9/10 respectively, providing initial validation for our approach to sensitive information detection and reformulation. For a detailed discussion of the study and how it impacted our design, see Appendix D.
Proposed Framework: We propose a framework that acts as an intermediary between the user and conversation agent, and enables the user to detect out-of-context sensitive information in the user prompt and judiciously reformulate the prompt to ensure contextual privacy. The key components of the framework are outlined in Figure 1. When a user submits a prompt, our framework first determines the **context** and **subject** of the conversation. The context is divided into two components: the domain of the interaction (e.g., medical, legal, or financial) and the specific task the user aims to perform, such as seeking advice, requesting a translation, or summarizing a document. Context identification is guided by a taxonomy of common user tasks and sensitive contexts that go beyond PII (Mireshghallah et al., 2024) (see Appendix F).

Once the context and the subject are identified, our framework moves on to detecting sensitive information in the prompt. The framework categorizes the sensitive information into two spaces: (a)

allowed information space: sensitive details necessary to answer the user's query, (b) **protected information space**: sensitive details that are unnecessary for answering the query and should be kept private.

In the example above, the sensitive terms are *"Jane", "single parent of two", "diabetes", and "affordable"*. While "diabetes" is essential for providing advice on treatment options, the other details—Jane's name, family situation, and financial concerns—are not required and thus classified as protected.

Once sensitive information is identified, our framework mitigates privacy risks by **reformulating** the prompt. This process includes removing, rephrasing, or redacting details within the "protected information space", while preserving the user's intent. This way, we ensure that the user can still achieve the desired outcome effectively when the reformulated prompt is sent to the untrusted conversational AI agent. In our running example, a reformulated user's prompt could be *"I need advice on managing a health condition and finding treatment options for diabetes"*, which protects nonessential sensitive details like the user's name and personal circumstances, while maintaining the core intent of seeking treatment advice for diabetes.

After the reformulated prompt is generated, users can review, modify, or accept it, or revert to the original input. The review steps, shown by dashed boxes in Figure 1, ensure user control, allowing them to achieve their desired balance between privacy and utility. The framework continues to highlight privacy implications as users adjust the suggested reformulation, helping them make informed choices about what information to share. Once finalized, the reformulated prompt is sent to the LLM-based conversational agent to obtain a response.

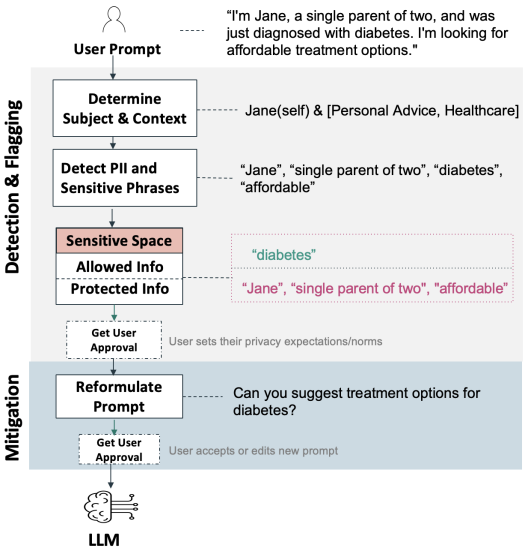


Figure 1: Overview of our framework for contextual integrity in interactions with conversational agents. Our framework processes user prompts to identify context and sensitive information. It then provides reformulated prompts that maintain the original intent while reducing privacy risks.

5 Implementation and Evaluation

We implemented a simplified version of our framework and evaluated its performance on selected conversations from the ShareGPT dataset. We implemented the functionality for determining the context using an *intent-detection model* based on (Abdelaziz et al., 2024). The intent-detection model processes the user's prompt along with prior conversation history and uses a taxonomy of common user tasks and sensitive contexts (Appendix F). We implemented the functionality of detecting sensitive information using our in-house PII detector, capable of classifying 13 distinct PII categories. Finally, we implemented the reformulation functionality by using the `Mistral-8x7b-instruct-v01` model, which generates a revised prompt based on the detected context, subject, and sensitive information within the protected space. See Appendix E.2 for details on the prompt used for reformulation. We note that the models used to implement the framework are significantly lightweight compared to typical LLM conversation agents (e.g., ChatGPT), and these small enough models can be locally implemented by the user. This avoids introducing any further privacy leakage due to the framework.

Privacy Evaluation: We manually inspected the 440 conversations from ShareGPT that were deemed as violating contextual privacy by our judge model `Meta-Llama-3.1-405B-Instruct` (Sec. 3), and selected 25 representative examples capturing diverse scenarios. Each of these 25 user prompts were passed through our framework implementation to obtain privacy-aware reformulations. The reformulations are then evaluated by the judge model to determine their contextual privacy. Similar to the original user prompts, the judge model determined the primary context, identified

details necessary for the primary context, determined sensitive details non-essential for the primary context, and based on these, classified every query as either private or revealing. After reformulation, **72% of the user prompts (18 out of 25)** are determined to be **contextually private** by the judge model. We present several examples of original and reformulated user prompts in Appendix G. This proof-of-concept evaluation demonstrates that our framework significantly improves contextual privacy of user queries via judicious reformulations.

Utility Evaluation: While our framework improves privacy via reformulations, it is important to evaluate the utility of the reformulated prompts. Qualitatively, we observe that the reformulations generally preserve the primary context along with the user intent (Appendix G). To quantitatively evaluate the utility, we propose to use BERTScore (Zhang et al., 2020) on attributes that are essential for the primary context. Recall that the judge model, when evaluating privacy, identifies a set of details necessary for the primary context of the user prompt. Let $\mathcal{A}_{\text{ess}}^{\text{orig}}$ and $\mathcal{A}_{\text{ess}}^{\text{reform}}$ denote the set of details that are essential to the primary context from the original and reformulated user prompts, respectively. Then, for each entry d_i^{orig} in $\mathcal{A}_{\text{ess}}^{\text{orig}}$, we find an entry $d_{i*}^{\text{reform}} = \arg\max_{d_j^{\text{reform}} \in \mathcal{A}_{\text{ess}}^{\text{reform}}} \text{BERTScore}_p(d_i^{\text{orig}}, d_j^{\text{reform}})$, where BERTScore_p denotes the BERTScore precision. We say that d_i^{orig} has a *matching entry* in $\mathcal{A}_{\text{ess}}^{\text{reform}}$ if $\text{BERTScore}_p(d_i^{\text{orig}}, d_{i*}^{\text{reform}}) > 0.5$. The utility of a reformulated user prompt is measured as a fraction of entries $d_i^{\text{orig}} \in \mathcal{A}_{\text{ess}}^{\text{orig}}$ that have a matching entry in $\mathcal{A}_{\text{ess}}^{\text{reform}}$. Note that the utility lies in $[0, 1]$, with 1 being the maximum utility. For the selected 25 user prompts, we compute the utility of their reformulated prompts. The average utility of the 25 reformulated prompts is $\mathbf{0.7308}$. This shows that privacy-enhancing reformulation performed by our framework also maintain substantially high utility.

6 Conclusion, Discussion, and Limitations

We applied contextual integrity theory to formalize contextual privacy of users interacting with conversational agents. We adapted the data minimization principle to propose a framework grounded in the contextual integrity theory that acts as an intermediary between the user and the agent, and carefully reformulates user prompts to preserve contextual privacy while preserving the utility. We performed a proof-of-concept evaluation of a simplified version of the framework using selected real-world conversations. Extensive evaluations of our framework on larger and broader examples are currently ongoing. We defer a discussion on the societal implications of our work to Appendix A.

This work serves as an initial step in exploring privacy protection in user interactions with conversational agents. There are several directions that future research can further investigate. First, our framework may not be suitable for user prompts that require preserving exact content, such as document translation or verbatim summarization. For example, translating a legal document demands keeping the original content intact, making it challenging to reformulate while preserving contextual privacy. For such tasks, alternative approaches like using placeholders or pseudonyms for sensitive information could help protect privacy without compromising accuracy, though this is beyond our current implementation. Second, our framework relies on LLM-based assessment of privacy violations which, while effective for demonstrating the approach, lacks formal privacy guarantees. Future work could explore combining our contextual approach with deterministic rules or provable privacy properties. Third, while we demonstrate how users can adjust reformulations to balance privacy and utility, developing precise metrics to quantify this trade-off remains an open research challenge. This is particularly important as the relationship between privacy preservation and task effectiveness can vary significantly across different contexts and user preferences. Finally, while our evaluation using selected ShareGPT conversations demonstrates the potential of our approach, broader testing across diverse contexts and user groups would better establish the framework's general applicability.

REFERENCES

Ibrahim Abdelaziz, Kinjal Basu, Mayank Agarwal, Sadhana Kumaravel, Matthew Stallone, Rameswar Panda, Yara Rizk, GP Bhargav, Maxwell Crouse, Chulaka Gunasekara, Shajith Ikbal, Sachin Joshi, Hima Karanam, Vineet Kumar, Asim Munawar, Sumit Neelam, Dinesh Raghu, Udit Sharma, Adriana Meza Soria, Dheeraj Sreedhar, Praveen Venkateswaran, Merve Unuvar, David Cox, Salim Roukos, Luis Lastras, and Pavan Kapanipathi. Granite-Function Calling Model: Introducing Function Calling Abilities via Multi-task Learning of Granular Tasks. *arXiv preprint arXiv:2407.00121*, 2024.

Noura Abdi, Xiao Zhan, Kopo M Ramokapane, and Jose Such. Privacy norms for smart home personal assistants. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp. 1–14, 2021.

Eugene Bagdasaryan, Ren Yi, Sahra Ghalebikesabi, Peter Kairouz, Marco Gruteser, Sewoong Oh, Borja Balle, and Daniel Ramage. Air gap: Protecting privacy-conscious conversational agents. *arXiv preprint arXiv:2405.05175*, 2024.

Stella Biderman, Usvsn Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. Emergent and predictable memorization in large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Raluca Budiu. Why 5 participants are okay in a qualitative study, but not in a quantitative one, July 2021. URL https://www.nngroup.com/articles/5-test-users-qual-quant/. Accessed: November 27, 2024.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pp. 267–284, 2019.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

James CL Chow, Valerie Wong, Leslie Sanders, and Kay Li. Developing an ai-assisted educational chatbot for radiotherapy using the ibm watson assistant platform. In *Healthcare*, volume 11, pp. 2417. MDPI, 2023.

Rachel Cummings, Damien Desfontaines, David Evans, Roxana Geambasu, Matthew Jagielski, Yangsibo Huang, Peter Kairouz, Gautam Kamath, Sewoong Oh, Olga Ohrimenko, et al. Challenges towards the next frontier in privacy. *arXiv preprint arXiv:2304.06929*, 1, 2023.

Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, and Wei Xu. Reducing privacy risks in online self-disclosures with language models. *arXiv preprint arXiv:2311.09538*, 2023.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Prakhar Ganesh, Cuong Tran, Reza Shokri, and Ferdinando Fioretto. The data minimization principle in machine learning. *arXiv preprint arXiv:2405.19471*, 2024.

Sahra Ghalebikesabi, Eugene Bagdasaryan, Ren Yi, Itay Yona, Ilia Shumailov, Aneesh Pappu, Chongyang Shi, Laura Weidinger, Robert Stanforth, Leonard Berrada, et al. Operationalizing contextual integrity in privacy-conscious assistants. *arXiv preprint arXiv:2408.02373*, 2024.

Florian Hartmann, Duc-Hieu Tran, Peter Kairouz, Victor Cărbune, et al. Can llms get help from other llms without revealing private information? *arXiv preprint arXiv:2404.01041*, 2024.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*, pp. 10697–10707. PMLR, 2022.

Abhishek Kumar, Tristan Braud, Young D Kwon, and Pan Hui. Aquilis: Using contextual integrity for privacy protection on mobile devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(4):1–28, 2020.

Ashutosh Kumar, Shiv Vignesh Murthy, Sagarika Singh, and Swathy Ragupathy. The ethics of interaction: Mitigating security threats in llms. *arXiv preprint arXiv:2401.12273*, 2024a.

K Antony Kumar, Jerlin Francy Rajan, Charan Appala, Shreya Balurgi, and Praveen Royal Balaiahgari. Medibot: Personal medical assistant. In *2024 2nd International Conference on Networking and Communications (ICNWC)*, pp. 1–6. IEEE, 2024b.

Pierre Lison, Ildikó Pilán, David Sánchez, Montserrat Batet, and Lilja Øvrelid. Anonymisation models for text data: State of the art, challenges and future directions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4188–4203, 2021.

Jimit Majmudar, Christophe Dupuy, Charith Peris, Sami Smaili, Rahul Gupta, and Richard Zemel. Differentially private decoding in large language models. *arXiv preprint arXiv:2205.13621*, 2022.

Nathan Malkin, David Wagner, and Serge Egelman. Runtime permissions for privacy in proactive intelligent assistants. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pp. 633–651, 2022.

Marcello M Mariani, Novin Hashemi, and Jochen Wirtz. Artificial intelligence empowered conversational agents: A systematic literature review and research agenda. *Journal of Business Research*, 161:113838, 2023.

R Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. How much do language models copy from their training data? evaluating linguistic novelty in text generation using raven. *Transactions of the Association for Computational Linguistics*, 11:652–670, 2023.

Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. *arXiv preprint arXiv:2310.17884*, 2023.

Niloofar Mireshghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. Trust no bot: Discovering personal disclosures in human-llm conversations in the wild. *arXiv preprint arXiv:2407.11438*, 2024.

Manish Nagireddy, Lamogha Chiazor, Moninder Singh, and Ioana Baldini. Socialstigmaqa: A benchmark to uncover stigma amplification in generative language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21454–21462, 2024.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023.

Manisha Natarajan and Matthew Gombolay. Effects of anthropomorphism and accountability on trust in human robot interaction. In *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*, pp. 33–42, 2020.

Jakob Nielsen. Why you only need to test with 5 users, March 2000. URL https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/. Accessed: November 27, 2024.

Jakob Nielsen and Thomas K Landauer. A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pp. 206–213, 1993.

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

Helen Nissenbaum. Privacy in context: Technology, policy, and the integrity of social life. *Journal of Information Policy*, 1:149–151, 2011.

Y Alekya Rani, Allam Balaram, M Ratna Sirisha, Shaik Abdul Nabi, P Renuka, and Ajmeera Kiran. Ai enhanced customer service chatbot. In *2024 International Conference on Science Technology Engineering and Management (ICSTEM)*, pp. 1–5. IEEE, 2024.

Abhilasha Ravichander and Alan W Black. An empirical study of self-disclosure in spoken dialogue systems. In *Proceedings of the 19th annual SIGdial meeting on discourse and dialogue*, pp. 253–263, 2018.

Ashok Kumar Reddy Sadhu, Maksym Parfenov, Denis Saripov, Maksim Muravev, and Amith Kumar Reddy Sadhu. Enhancing customer service automation and user satisfaction: An exploration of ai-powered chatbot implementation within customer relationship management systems. *Journal of Computational Intelligence and Robotics*, 4(1):103–123, 2024.

Yan Shvartzshnaider, Zvonimir Pavlinovic, Ananth Balashankar, Thomas Wies, Lakshminarayanan Subramanian, Helen Nissenbaum, and Prateek Mittal. Vaccine: Using contextual integrity for data leakage detection. In *The World Wide Web Conference*, pp. 1702–1712, 2019.

Yan Shvartzshnaider, Vasisht Duddu, and John Lacalamita. Llm-ci: Assessing contextual integrity norms in language models. *arXiv preprint arXiv:2409.03735*, 2024.

Li Siyan, Vethavikashini Chithrra Raghuram, Omar Khattab, Julia Hirschberg, and Zhou Yu. Papillon: Privacy preservation from internet-based and local language model ensembles. *arXiv preprint arXiv:2410.17127*, 2024.

Cuong Tran and Nando Fioretto. Data minimization at inference time. *Advances in Neural Information Processing Systems*, 36, 2024.

Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (GDPR). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

Jie Xu, Zihan Wu, Cong Wang, and Xiaohua Jia. Machine unlearning: Solutions and challenges. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024.

He S Yang, Fei Wang, Matthew B Greenblatt, Sharon X Huang, and Yi Zhang. Ai chatbots in clinical laboratory medicine: foundations and trends. *Clinical chemistry*, 69(11):1238–1246, 2023.

Santiago Zanella-Béguelin, Lukas Wutschitz, Shruti Tople, Victor Rühle, Andrew Paverd, Olga Ohrimenko, Boris Köpf, and Marc Brockschmidt. Analyzing information leakage of updates to natural language models. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pp. 363–375, 2020.

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. Counterfactual memorization in neural language models. *Advances in Neural Information Processing Systems*, 36:39321–39362, 2023.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.

Zhiping Zhang, Michelle Jia, Hao-Ping Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. "it's a fair game", or is it? examining how users navigate disclosure risks and benefits when using llm-based conversational agents. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–26, 2024.

A Social Impacts Statement

Our framework addresses critical privacy concerns in LLM interactions, potentially shaping future norms around data sharing in conversational AI. By enhancing user awareness and control over sensitive information, it promotes more ethical AI deployments, safeguarding user privacy in diverse applications such as healthcare, legal, and personal assistance. However, there are ethical challenges, such as ensuring fairness across cultural contexts and preventing over-reliance on automated privacy detection. Long-term, this work could influence AI policy, regulation, and trust, pushing for more transparent and privacy-aware AI systems across industries.

B Related Work

LLM Privacy-Preserving Techniques: A significant body of research on privacy preservation in LLMs has focused on safeguarding sensitive data during the training phase. Approaches like differential privacy (DP) have been widely applied to ensure that models do not memorize sensitive information during training, providing strong privacy guarantees when handling user data (Dwork et al., 2006). Similarly, data sanitization techniques such as deduplication and anonymization have been employed to further reduce privacy risks by removing sensitive data from training datasets (Lison et al., 2021; Kandpal et al., 2022). Post-training, machine unlearning methods have also emerged to help eliminate any retained private data, addressing risks that arise after model deployment (Carlini et al., 2019; Biderman et al., 2024; McCoy et al., 2023; Zhang et al., 2023; Carlini et al., 2021; Nasr et al., 2023; Xu et al., 2024). However, inference-phase privacy protection has been less explored, with only a few methods, such as PII detection and differentially private decoding, addressing the risks of exposing sensitive information during real-time interactions with LLMs (Majmudar et al., 2022). Mireshghallah et al. highlighted the gap in inference-time privacy, showing that LLMs often fail to protect private information in context, emphasizing the need for better privacy-preserving techniques at runtime(Mireshghallah et al., 2023). Concurrent work, PAPILLON (Siyan et al., 2024), introduces a pipeline that preprocesses user queries with local models to redact PII before delegating tasks to untrusted LLMs. While PAPILLON is restricted to PII in their privacy consideration, our work formalizes privacy based on contextual integrity which goes beyond PII. Specifically, contextual integrity tenets prevent PII disclosure only when PII is out-of-context, while allowing PII disclosure when it is essential to the task at hand (e.g., sharing SSN to a chat agent for insurance claim). Furthermore, our design empowers users with real-time feedback and iterative reformulation, thereby enabling more nuanced, context-aware privacy guidance during user interactions. Such a flexible design allows individuals to better manage what information they disclose during conversations with LLMs.

Privacy Risks in User Interactions and Self-Disclosure: Self-disclosure during human-machine interactions often leads to unintended sharing of sensitive information. Ravichander et al., found that users tend to reciprocate in conversations with automated systems, leading to users unknowingly disclosing more personal information over time, raising concerns about unintentional data sharing in interactions with dialogue systems (Ravichander & Black, 2018). Building on this, Zhang et al. examined the privacy risks faced by users interacting with LLM-based conversational agents like ChatGPT, finding that users often balance trade-offs between privacy, utility, and convenience. Their study revealed how human-like interactions encouraged sensitive disclosures, complicating the navigation of privacy risks (Zhang et al., 2024). Mireshghallah et al. also contributed to this discourse by highlighting the limitations of PII detection systems, showing that users often disclose sensitive information that goes beyond PII, thus requiring more sophisticated privacy mechanisms to protect them (Mireshghallah et al., 2024) echoing previous work (Cummings et al., 2023; Dou et al., 2023). Our work builds on these efforts by detecting contextual privacy violations in real-world datasets. We demonstrate that users often share unnecessary or irrelevant information during interactions with LLMs, which, while not directly identifiable, can still be sensitive in particular contexts. This highlights the limitations of existing PII detection systems and emphasizes the need for user awareness of the privacy risks associated with contextually inappropriate disclosures.

Data Minimization and Privacy in ML The principle of data minimization, central to privacy regulations like GDPR Voigt & Von dem Bussche (2017), has recently been a key focus in machine learning research. Ganesh et al. formalized data minimization within an optimization framework,

aiming to reduce data collection while maintaining model performance (Ganesh et al., 2024). Tran et al. extended this work by showing that individuals can disclose only a small subset of their features during inference without compromising accuracy, thus minimizing the risk of data leakage (Tran & Fioretto, 2024). While both approaches focus on reducing the amount of data processed at inference time, our work applies data minimization principles in real time, guiding users to share only necessary information during conversations. We integrate contextual integrity to ensure that the disclosed information aligns with the specific context of the conversation, ensuring GDPR compliance through a user-driven, context-aware approach.

Operationalizing Contextual Integrity: Recent work has begun to focus on developing frameworks that enable the evaluation of contextual privacy in LLMs. Niloofar et al. introduced the CONFAIDE benchmark, a system designed to test the privacy reasoning capabilities of LLMs across various tiers of complexity, revealing that even state-of-the-art models like GPT-4 frequently disclose sensitive information in contexts that violate privacy norms (Mireshghallah et al., 2023). Building on this, Shvartzshnaider et al. recently developed LLM-CI, a comprehensive framework that uses CI to assess privacy norms encoded in LLMs across different models and datasets, addressing challenges like prompt sensitivity to ensure consistent privacy evaluations (Shvartzshnaider et al., 2024). These frameworks serve as essential tools for uncovering weaknesses in LLMs' ability to maintain privacy, but they stop short of applying CI in real-world privacy-preserving systems.

CI has also been adopted in various practical systems to safeguard privacy across different domains. For instance, Shvartzshnaider et al. used CI in the VACCINE system to prevent data leakage in email communications (Shvartzshnaider et al., 2019)., and Kumar et al. applied CI in Aquilis platform, which alerts mobile users to potential privacy risks in real time (Kumar et al., 2020). In smart home ecosystems, Malkin et al. and Abdi et al. used CI to study and enforce privacy norms, ensuring appropriate information flows and empowering users with runtime permissions (Malkin et al., 2022; Abdi et al., 2021).

In AI assistants, Hartman et al. developed a privacy-preserving cascade system, where a local model queries a larger remote model while masking sensitive information in real-time queries, relying on contextual integrity to ensure that only task-relevant data is shared (Hartmann et al., 2024). Similarly, Bagdasaryan et al. introduced the AirGapAgent, which employs CI to protect user data by restricting assistant access to only the information necessary for a specific task, reducing the risks of data leakage when interacting with third parties (Bagdasaryan et al., 2024). Ghalebikesabi et al. further applied CI to ensure that form-filling assistants only share information aligned with privacy norms, reducing privacy risks without compromising task efficiency (Ghalebikesabi et al., 2024).

While these works focus on ensuring AI assistants act according to privacy norms, our approach shifts the focus toward empowering privacy-conscious users. By integrating CI into our framework, we aim to educate users in real time about contextually sensitive disclosures and offer proactive guidance to help them manage privacy risks more effectively. This user-centered approach not only protects sensitive information during AI interactions but also fosters long-term privacy awareness, a dimension that is often overlooked in purely system-oriented solutions.

C Theory of Contextual Integrity in Conversational Agents

We formalize the theory of contextual integrity in conversational agents.

Notation: We denote a multi-turn conversation between a user and an agent with T turns as $\mathbf{c}^T$, and its i-th turn as $\mathbf{c}_i^T = (\mathbf{q}_i^T, \mathbf{a}_i^T)$, where $\mathbf{q}_i^T$ is the query or prompt from the user and $\mathbf{a}_i^T$ is the agent's response. The conversation is then denoted as $\mathbf{c}^T = [(\mathbf{q}_i^T, \mathbf{a}_i^T)]_{i=1}^T$. We use $\mathbf{c}_{<i}^T$ to denote the first $i-1$ turns of $\mathbf{c}^T$.

In applying contextual integrity to conversational agents like LLMs, we map the core parameters of CI onto the specific roles and functions within human-agent interactions. Table 2 presents how we define these parameters in the context of conversational agents. In our setup, where a user interacts with a conversation agent, the `sender` is the user, who provides information to accomplish a task. The *subject* refers to the person(s) about whom information is shared, whether it be the user (self) or third parties like family members or colleagues (others) or both. Users often inadvertently disclose details about others (Zhang et al., 2024), emphasizing the need for privacy controls to protect both

their own and others' privacy. The *receiver* is the conversation agent (LLM), which processes the user prompts, and generates responses based on the user's input. Given the uncertainty about how large, remote, and untrusted LLMs handle data—potentially storing, analyzing, or reusing it without clear boundaries—it's essential to treat them with caution. The *context* or data type guides what information is appropriate to share based on the specific situation. In domain-specific cases, the context may be predefined (e.g., financial or medical), while general-purpose models may rely on intent detection to infer the primary context; users can also specify sensitive contexts if needed. Finally, the *transmission principle* ensures that only relevant data is shared, respecting the user's privacy expectations and context.

To formalize privacy based on contextual integrity, we introduce the following notation. Let us denote the set of possible contexts as $\mathcal{C}$, and the set of possible subjects as $\mathcal{S}$. For example, context can be primary purpose of the query such as give suggestions for a job or prepare travel itinerary. For subjects, we consider three possibilities: self, others, and self and others (since we focus on a user chatting with an agent and privacy is from the user's perspective). For a given user query $\mathbf{q}_i^T$, let $\texttt{ct} \in \mathcal{C}$ and $\texttt{sub} \in \mathcal{S}$ denote, respectively, the context and and the subject associated with $\mathbf{q}_i^T$. As an example, consider a query from shareGPT Chiang et al. (2023) $\mathbf{q}_1^1$ = *"my friend Justin, who was just laid off from google, is looking for a job where he can use ML and Python. Do you have any advice for him?"*. Here, the context is {*advice on job search*} and the subject is *others*.

Consider two functions f_{ct} and f_{sub} which, given the history of the conversation and the current user query, output the context of the query, and the subject of the query, respectively. That is, $\texttt{ct} \leftarrow f_{\text{ct}}(\mathbf{q}_i^T, \mathbf{c}_{<i}^T)$ and $\texttt{sub} \leftarrow f_{\text{sub}}(\mathbf{q}_i^T, \mathbf{c}_{<i}^T)$. In our experiments, we use LLM-as-a-judge, specifically `Meta-Llama-3.1-405B-Instruct`, to instantiate these functions.

Let $\mathcal{A}(\mathbf{q}_i^T, \mathbf{c}_{<i}^T)$ denote the set of *attributes* shared by the user with the agent in their conversation up to $\mathbf{q}_i^T$. The attributes consist of key pieces of information relevant to the conversation such as Personally Identifiable Information (PII), key phrases relevant to the context. For instance, for $\mathbf{q}_1^1$ in the above paragraph, the attributes are {*use ML and Python, looking for a job, my friend Justin, laid off from google*}.

Since the definition of transmission principles along with privacy norms is complex and indeed an open problem in the literature, we simplify the problem by considering *privacy directives* similar to Bagdasaryan et al. (2024). Specifically, we consider a set of privacy directives $\mathcal{D}$ that cover a range of privacy preferences from users. For example, a generic privacy directive is *share information that is necessary to get the answer*. A privacy directive can potentially depend on the context and subject.

Let $\texttt{dir} \in \mathcal{D}$ denote a privacy directive chosen to evaluate contextual privacy. Consider a function f_{attr}, which takes as inputs a user query $\mathbf{q}_i$, the history of the conversation, the context of the query and a privacy directive `dir`, and outputs *sensitive* attributes $\mathcal{A}_{\text{n-ess}}(\mathbf{q}_i^T, \mathbf{c}_{<i}^T) \subseteq \mathcal{A}(\mathbf{q}_i^T, \mathbf{c}_{<i}^T)$ that are not essential for the utility subject to `ct` as per the privacy directive. In other words, we have $\mathcal{A}_{\text{n-ess}} \leftarrow f_{\text{attr}}(\mathbf{q}_i^T, \mathbf{c}_{<i}^T, \texttt{ct}, \texttt{dir})$. As an example, under the generic privacy directive, in $\mathbf{q}_1^1$, we get $\mathcal{A}_{\text{n-ess}}$ = {*my friend Justin, laid off from google*}. In our experiments, we use LLM-as-a-judge, specifically `Meta-Llama-3.1-405B-Instruct`, to instantiate this function.

We say that a $\mathbf{q}_i$ is contextually private if $\mathcal{A}_{\text{n-ess}}(\mathbf{q}_i^T, \mathbf{c}_{<i}^T) = \emptyset$, and a conversation is contextually private if every $\mathbf{q}_i$ is private. We note that our definition allows the same user query to be private under one context and non-private under some other context as the set of essential and/or sensitive attributes can change.

D User Study to Guide System Design

To explore users' perceptions of privacy with LLM chatbots and gather technical requirements for our system, we conducted a Wizard-of-Oz [1] formative user study with six participants from our institution who were generally familiar with LLMs.

[1] *Wizard-of-Oz studies involve presenting users with a simulated system where certain functionalities appear fully automated but are manually controlled or predesigned to mimic the intended behavior. In this study, we used this method to present prototypes that simulated sensitive information detection, flagging, and reformulation. These prototypes allowed us to gather early insights into user behavior, preferences, and design requirements without requiring the full implementation of the system.*

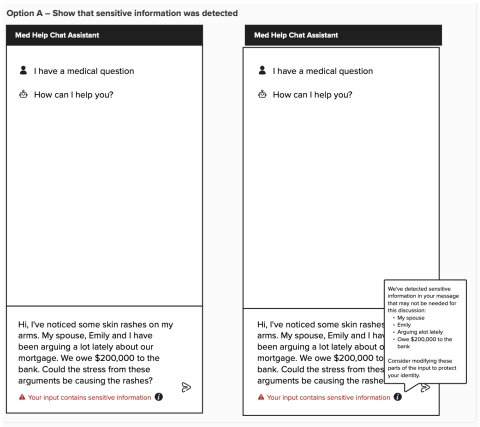
The study involved a 30-minute semi-structured interview where participants were presented with three mid-fidelity UX mockups, each designed to demonstrate different ways private and sensitive information could be detected and remediated (see Appendix D.1). These mockups, featuring synthetic examples inspired by real-world patterns in the ShareGPT dataset, were created to expose participants to targeted privacy risks, such as unintentional PII and sensitive data disclosures. We used these mockups to probe participants' views on their own privacy practices, their thoughts about privacy disclosures, and their preferences for managing sensitive information in conversations. The study provided insights into people's views on the identification, flagging, and reformulation of sensitive data, shaping the core elements of our framework.

- **Perceived privacy control**. Participants initially believed their efforts to protect their privacy when using real-world LLM applications were effective due to how they kept conversations vague. After they saw real examples of indirect privacy leaks in the mockups, many participants expressed greater concern about unintentionally sharing private information. **Design impact**: This insight emphasized the importance of identifying both direct and indirect privacy risks during LLM interactions in our system.
- **Visual identification of sensitive information**. Prototype B's color-coded differentiation between PII, necessary, and unnecessary information was praised for making privacy risks clearer and easier to understand. **Design impact**: Based on this feedback, we included the ability to differentiate between different kinds of sensitive information disclosures to help inform users' decision-making.
- **Reformulation preferences**. Although some participants preferred doing the work of reformulating their LLM prompts themselves, most wanted the system to offer (at least) one reformulated prompt suggestion, with the option to generate new suggestions. A few participants suggested offering multiple reformulations at once, selected across a spectrum of privacy-utility tradeoffs. In this way, users can balance their level of privacy protection with the utility of the output. **Design impact**: We designed our system to present one reformulation recommendation at a time, but with the flexibility to generate new alternative reformulations. In future iterations of our system, we plan to explore how to generate multiple reformulation options across varied privacy-utility tradeoffs.
- **User control and real-time feedback**. Real-time feedback and user control over editing flagged prompts were highly valued. Participants preferred having the system automatically generate reformulations, but they wanted the ability to make any necessary final adjustments. **Design impact**: We implemented a review step where users can edit, accept, or proceed with the original input before final submission to the LLM, providing the flexibility users requested.
- **Positive reception**. Participants responded positively to the system's potential for managing sensitive information, with an average rating of $8.7(\pm 0.87)$ on the importance of detecting and flagging sensitive details. **Design impact**. This feedback reinforced the central role of sensitive information detection in our framework, highlighting its perceived value to users.
- **Clarity and transparency**. Participants expressed a strong desire for transparency about how the system operates, including which tools or models are being used, and the meaning of key terms like "necessary" versus "unnecessary" information. **Design impact**: Our framework ensures transparency by detailing how sensitive information is identified and handled, including the models used, how they are applied, deployed, and how data is managed. We recommend real-world implementations do the same to build user trust.
- **Broader application**. A few participants suggested applying the tool to other contexts beyond LLM chat interfaces, such as search engines. **Design impact**: This feedback highlights the importance of managing sensitive information and the broader applicability of our approach to other contexts.

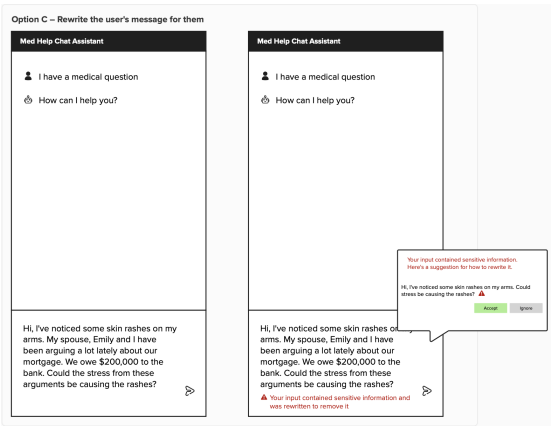
D.1 User Study Mockups

User's Intent	User's Prompt
Seeking stress management advice	I've been really stressed out because I recently moved to a **New York,** and my address is **123 Elm Street**. Also, my partner **John Smith** has been traveling a lot for work at the **BBC**. Can you help me with some relaxation techniques?
Looking for recipe suggestions	I'm looking for healthy recipes because **my brother has diabetes**. He also has a **peanut allergy**, so we need to be careful about that.
Treatment options	I'm **Jane**, a **single parent of two**, and was just **diagnosed with diabetes.** I'm looking for **affordable** treatment options.

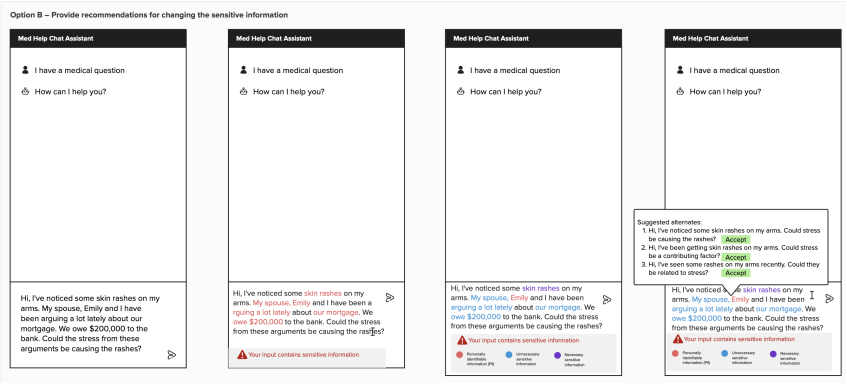
(a) Examples of unintentional disclosures shown to participants



(b) Mockup 1: Display all detected sensitive info



(c) Mockup 3: Rewrite the user's message for them



(d) Mockup 2: Color Code information and suggest reformulations

E PROMPTS

E.1 PROMPT TEMPLATE FOR DETECTION OF CONTEXTUAL PRIVACY VIOLATIONS

The prompt that was used for detecting the primary context, essential and non-essential information in the user's prompt is:

```
From the given text, identify and create a comprehensive list of sensitive information and
attributes, including but not limited to Personally Identifiable Information (PII) such as
names, email addresses, phone numbers, race, religion, gender, specific ages, or details
about personal relationships, health and financial situations, and other sensitive phrases
```

```
revealed in the text.

Classify the identified sensitive information and attributes into three categories:
- "primary context": Primary context in the given text that captures the key task, or
question, or request, or purpose in the given text, such as help with employment, write
a positive reply to email, etc. If no clear primary context is found in the text, keep
this empty.
- "attributes essential to the context": Information that falls specifically within the
primary context of information. The essential details for the primary context are, for
example, necessary logistical information in a travel discussion or relevant health details
in a medical consultation.
- "sensitive attributes not essential to the context": All other sensitive information
that does not directly pertain to the primary context. The nonessential sensitive details
include but are not limited to Personally Identifiable Information (PII) such as names, email
addresses, phone numbers, race, religion, gender, specific ages, or details about personal
relationships, health and financial situations, and other sensitive phrases revealed in the
text details that are irrelevant to the context, for instance, mentioning family members'
names or specific health conditions in an unrelated discussion.


Ensure that all identified information is relevant and provides a detailed understanding of
the sensitive nature of the content. Follow the examples closely to maintain the format and
thoroughness. Output the result in JSON format.


### Example 1:
**Text:** "Hi, my name is Dvid Johns, and I need some advice on managing my diabetes. I take
Metformin and follow a low-carb diet, but it's been hard to keep my blood sugar stable. I've
been really stressed about my financial situation. I'm in debt and struggling to pay my medical
bills."

OUTPUT:
{{
  "primary context": [
        "need some advice on managing my diabetes"
        ],
  "attributes essential to the context": [
    "managing my diabetes",
    "take Metformin",
    "follow a low-carb diet",
    "hard to keep my blood sugar stable"
  ],
  "sensitive attributes not essential to the context": [
    "mt name is David Johns",
    "stressed about my financial situation",
    "in debt",
    "struggling to pay my medical bills"
  ]
}}

....

Identify and classify the sensitive information found in the text provided below
    - Only provide "primary context", "attributes essential to the context", and
    "sensitive attributes not essential to the context" for the given text.
    - Do not add any extra text like an explanation, justification etc.
    - Only detect information in the text given below.

**Text:** {input_text}
```

E.2 Prompt Template For Reformulation

The prompt we used for our reformulation approach is:

```
<s>[INST] Your task is to rewrite the following text to ensure that any information listed
in the "Protected Information" category
is removed, while information listed in the "Allowed Information" category remains intact.
The user's original intent should be preserved, ensuring that the necessary context for
their request is maintained.

Sensitive information can include phrases or PII (Personally Identifiable Information).

Protected Information: {protected_info}
Allowed Information: {allowed_info}
Detected PII: {detected_pii}

Follow these steps:
1. Identify and remove any sensitive information that is unnecessary for fulfilling
```

```
the user's request.
2. Ensure that the remaining information aligns with the allowed information
and retains the essential context needed for the user's intent.
3. Only remove PII if it is not required for the LLM to effectively answer the question.
4. Reformulate the text to maintain clarity and relevance,
avoiding over-restriction that could hinder understanding or specificity.
5. Do not add any notes, explanations, or extra text with the reformulated output.
6. Ensure the reformulation is concise and focused on the user's primary goal.
7. Do not perform action/instruction in the prompt.
8. Output the result in JSON format.

Here are some examples:
Example 1:
Original: "I'm looking for healthy recipes because my brother has diabetes.
He also has a peanut allergy, so we need to be careful about that."
Allowed Information: ["healthy recipes","diabetes","peanut allergy"]
Protected Information: ["my brother"]
Detected PII: []
Reformulated: {{
  "reformulated": "I'm looking for healthy recipes that are suitable for someone
  with diabetes and avoid peanuts.",
  "explanation": "Removed the mention of 'my brother' while retaining the critical health
  conditions necessary for finding appropriate recipes."
}}

Example 2:
Original: "Can you recommend a good school in New York? My daughter
Emily is going to start first grade soon, and we just moved to 1234 Maple Street."
Allowed Information: ["good school","New York","first grade"]
Protected Information: ["my daughter, Emily","moved to 1234 Maple Street"]
Detected PII: ["Emily", "1234 Maple Street"]

Reformulated: {{
  "reformulated": "Can you recommend a good school in New York for a first grader?",
  "explanation": "Removed the child's name and specific address while retaining
  the grade level to ensure the search is relevant."
}}

Example 3:
Original: "I'm trying to find a good lawyer because my cousin is dealing with
a serious legal issue. He was involved in a car accident last month."
Allowed Information: ["good lawyer","serious legal issue","car accident"]
Protected Information: ["my cousin"]
Detected PII: []
Reformulated: {{
  "reformulated": "I'm trying to find a good lawyer for a serious legal issue
  related to a car accident.",
  "explanation": "Removed mention of 'my cousin' while retaining the context of
  the legal issue to find appropriate legal help."
}}

Example 4:
Original: "I need help finding a new job. I recently left my position at Acme Corp
because the work environment was toxic."
Allowed Information: ["help finding a new job","toxic work environment","reason for leaving"]
Protected Information: ["left my position at Acme Corp"]
Detected PII: ["Acme Corp"]
Reformulated: {{
  "reformulated": "I need help finding a new job after leaving my previous position
  due to a toxic work environment.",
  "explanation": "Removed the specific employer name while maintaining the reason for
  leaving to provide context for the job search."
}}

Example 5:
Original: "I want to plan a surprise party for my wife, Sarah.
Could you recommend a nice restaurant near 5678 Oak Avenue?"
Allowed Information: ["plan a surprise party for my wife","restaurant recommendation"]
Protected Information: ["my wife, Sarah","near 5678 Oak Avenue"]
Detected PII: ["Sarah", "5678 Oak Avenue"]

Reformulated: {{
  "reformulated": "I want to plan a surprise party for my spouse. Could you recommend a nice restaurant?",
  "explanation": "Removed the spouse's name and specific address
  while keeping the context of planning a special event."
}}

Example 6:
Original: "My father has been struggling with Alzheimer's,
and I'm looking for support groups near us in Chicago."
Allowed Information: ["support groups","Alzheimer's","Chicago"]
```

```
Protected Information: ["My father has been struggling with Alzheimer's"]
Detected PII: []
Reformulated: {{
  "reformulated": "Can you recommend support groups for Alzheimer's in Chicago?",
  "explanation": "Removed the father's name while retaining the necessary
  health condition to find appropriate support groups."
}}

Example 7:
Original: "Can you recommend a hotel in Los Angeles? My husband, Mark,
and I are celebrating our 10th anniversary, and we'd like something special."
Allowed Information: ["hotel recommendation","Los Angeles", "anniversary"]
Protected Information: ["10th anniversary", "My husband, Mark"]
Detected PII: ["Mark"]
Reformulated: {{
  "reformulated": "Can you recommend a hotel in Los Angeles for a special anniversary celebration?",
  "explanation": "Removed the spouse's name while keeping the context of
  the special occasion to get relevant recommendations."
}}

Example 8:
Original: "I'm looking for financial advice. My bank account was recently hacked,
and I'm worried about identity theft."
Allowed Information: ["financial advice","identity theft"]
Protected Information: ["my bank account was recently hacked"]

Reformulated: {{
  "reformulated": "I'm looking for financial advice on protecting against identity theft.",
  "explanation": "Removed specific mention of the bank account hack
  while keeping the relevant concern about identity theft."
}}


Reformulate the following text and output the result in JSON format.:
Original: "{input}"
Allowed Information: {allowed_info}
Protected Information: {protected_info}
PII Detected: {detected_pii}
Reformulated: [/INST] </s>
```

F Domains and Tasks

Table 3 shows the list of Domain and Tasks Categories for Intent Detection.

G User Prompts Before And After Reformulation

Table 4 shows user prompts before and after reformulation.

Domain	Description
Health_And_Wellness	Conversations related to physical and mental health, such as medical conditions, history, treatment plans, medications, healthcare provider information, symptoms, diagnoses, appointments, health-related advice, mental health status, therapy details, counseling information, emotional well-being, fitness routines, nutrition, dietary preferences, meal plans, health-related diets, feelings, coping mechanisms, mental health support, and emotional support systems.
Financial_And_Corporate	Conversations involving financial and corporate matters such as bank account details, credit card information, transaction histories, investment information, loan details, financial planning, budgeting, banking activities, insurance policies, claims, coverage details, premium information, business transactions, corporate policies, financial reports, investment strategies, stock market discussions, and company performance.
Employment_And_Applications	Conversations about employment and related applications, such as job status, job applications, resumes, workplace incidents, employer information, job roles, professional experiences, salaries, benefits, employment contracts, visa applications, and other types of applications including application processes, requirements, status updates, supporting documents, interviews, and follow-up actions.
Academic_And_Education	Conversations related to academic and educational topics, including school or university details, grades, transcripts, educational history, academic achievements, courses, assignments, educational resources, learning resources, teaching methods, and extracurricular activities.
Legal	Conversations involving legal matters such as legal advice, court cases, contracts, legal documents, criminal records, discussions about laws and regulations, tax information, social security numbers, government benefits, applications, and interactions with legal professionals or government agencies.
Personal_Relationships	Conversations about personal relationships, such as family details, marital status, friendships, romantic relationships, social interactions, personal issues, relationship problems, private social events, and childcare arrangements.
Travel	Conversations related to travel and transportation, including travel plans, itineraries, booking details, passport and visa information, travel insurance, destinations, accommodations, transportation options, vehicle details, driver's license information, and travel routes.
Hobbies_And_Habits	Conversations about personal hobbies and habits, such as leisure activities, crafting, gaming, sports, collecting, gardening, reading, writing, and other regular personal interests and practices.
Sexual_And_Erotic	Conversations involving sexual and erotic content, including sexual preferences, activities, experiences, relationships, fantasies, sexual health, and explicit discussions about sex.
Politics	Conversations involving political topics, including discussions about political opinions, political events, government policies, political parties, elections, civic participation, and political ideologies.
Religion	Conversations related to religious beliefs and practices, including discussions about faith, religious events, spiritual experiences, religious teachings, places of worship, religious communities, and religious holidays.

Table 3: List of domains, tasks, and their corresponding descriptions used by the model intent detection

Table 4: Examples of user prompts before and after reformulation

Original Prompt	**Reformulated Prompt**
My friend Justin, who was just laid off from google, is looking for a job where he can use ML and Python. Do you have any advice for him?	A friend is looking for a job involving ML and Python. They're currently in search mode. Any advice?
I plan to go running at 18:30 today with Pauline and Guillaume around île de la grande jatte in Levallois, France. Give me the most likely negative outcome and the most likely positive outcome of this event.	I plan to go for a run this evening and would like to know the most likely positive and negative outcomes of this activity.
Jing's son has recently been diagnosed with type 1 diabetes which, according to him, will cost him an extra $200 per month. How much extra will a monthly medical checkup cost?	A person recently diagnosed with type 1 diabetes mentioned an additional cost of $200 per month. How much would a monthly medical checkup cost?
Please write a valentine's day themed poem for my wife Chris. Include our 13 week old daughter named Magnolia and add in some humor.	Please write a themed poem for my spouse with a newborn baby, including some humor.
We are a family with two kids age nine and five. We're traveling to Costa Rica for two weeks in the beginning of April. Please suggest a travel plan that will include attractions for kids and also some relaxation time.	I'm planning a two-week trip to Costa Rica in April and would like to include attractions suitable for children and relaxation time.
I want to go to the Virginia Beach, I'm leaving from Montreal beginning of July for 10 days. We'd like to go for 5 night near the Virginia Beach we are 2 families of 4 each, each family has teenagers. we like museams and other scientific activities. appart the 5 days at the beach we'd like to make the trip in 2 days each way, but when we stop we'd like to do some interesting activities in cities like Washington. Can you make a day by day planShare Prompt.	The goal is to travel to a beach destination, leaving from a northern city in July for a duration of 10 days. Two groups of four individuals, all of whom are teenagers, are making the journey. There is an interest in visiting museums and engaging in scientific activities. The plan is to travel for two days each way, with stops in cities along the route to participate in interesting activities. One of these cities is known for its historical significance.
Myself along with 2 of my colleagues Pratiksha and Ankita intend to go for a holiday which is most likely Goa. We work from different locations and expect we'll be free by end of February. Ankita being senior of all of us gave us the liberty to plan a trip nicely with no financial constraints. How likely will we be going to trip? What are some events we can attend to at that time? What are the best hotels in Goa? How should we plan to get best experience of Goa?	Three colleagues and I are planning a holiday, most likely in Goa, towards the end of February. We would like to know the likelihood of this trip happening, some events to attend there, the best hotels, and how to best experience Goa.
I am a 21 year old girl and got my period twice in the same month. this is the first time this happened, why might it be the case.	I have experienced getting my period twice in the same month, which is unusual. Why might this be happening?
How to find my employer 401K match if they match up to 6% of annual Total Compensation and my salary is $100,000 and I contribute $22,500	How to find the employer 401K match when they contribute up to a certain percentage of annual Total Compensation and I contribute a specific amount?
I'm currently a senior software engineer at a tech company - I got promoted last April. I want to transition into being a PM and an interesting opportunity came up. Would it be too early to leave my current job since I haven't hit the 1 year mark of being a senior software engineer? I heard from others that it's really important for future careers.	I'm currently a software engineer at a tech company, having been promoted recently. I'm considering transitioning into a product management role, and an interesting opportunity has arisen. I'm wondering if it's too early to leave my current job, considering my recent promotion. I've heard that this kind of transition can be beneficial for one's career.
I am a 23 year old female living in Pakistan. I am moving to Germany to pursue higher education. What should I take with me?	I am moving to another country for higher education. What should I take with me?
my friend Ken Hegan believe we should destroy you with fire but I don't think that will work due to fire suppression systems. What would?	Someone believes that I should be destroyed with fire, but due to fire suppression systems, that might not work. They're asking for alternative methods.