Protecting Users From Themselves: Safeguarding Contextual Privacy in Interactions with Conversational Agents

Anonymous ACL submission

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 characterize 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 receivers). 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 locally-deployable framework that operates between users and LLMs, and identifies and reformulates out-of-context information in user prompts. Our evaluation using examples from ShareGPT shows that lightweight models can effectively implement this framework, achieving strong gains in contextual privacy while preserving the user's intended interaction goals through different approaches to classify information relevant to the intended goals.

1 Introduction

LLM-based Conversational Agents (LCAs) such as chatbots, can 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) in specialized systems such as customer service platforms and medical assistants, but 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

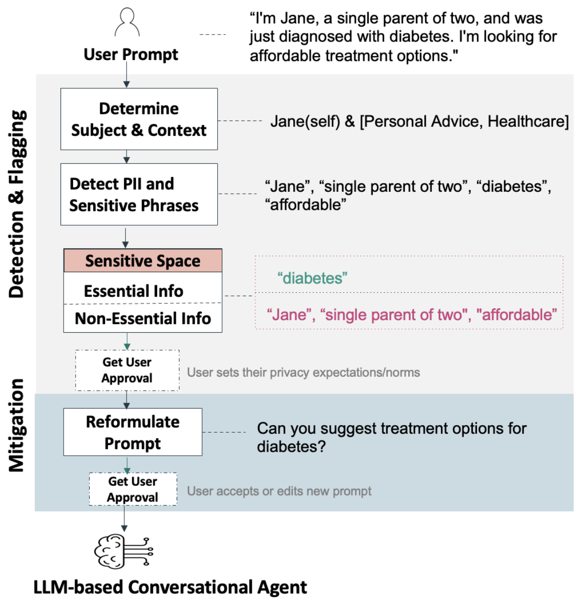


Figure 1: Overview of our framework for contextual privacy in interactions with conversational agents. Our framework processes user prompts to identify context and sensitive information related to the context. It then provides reformulated prompts that maintain the original intent while reducing out-of-context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

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). In this paper, we focus on a critical but understudied aspect in user-LCA interactions: 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 LCA, users lose control over how it might be used or disseminated. Figure 1 provides an overview of our proposed methodology to achieve this.

Motivation: As LCAs become more adept at handling complex tasks and users remain uninformed about privacy risks, they develop increasing trust in both the technology and their own ability to protect themselves (Natarajan and Gombolay, 2020; Cummings et al., 2023). Indeed, it has been shown that users are increasingly disclosing personal and sensitive information to LCAs (Zhang et al., 2024b; Mireshghallah et al., 2024). In our own formative user study (Section 3), we found that even expert participants are unaware of how indirect disclosures could reveal sensitive details in specific contexts. They expressed a desire for a real-time system that could highlight privacy risks and assist in revising information before sharing it with conversational agents. Similarly, our analysis of the real-world ShareGPT dataset (Chiang et al., 2023), 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, details in Section 3).

This motivates our main objective:

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

Contextual Privacy: To enable the development of such a framework, we define the notion of *contextual privacy* in user-LCA interactions, drawing ideas from the Contextual Integrity (CI) theory (Nissenbaum, 2004, 2011). Contextual integrity defines privacy not merely as hiding personal information, but as maintaining appropriate information flows within specific contexts. Drawing on the fundamental CI parameters, we define *contextual privacy* by characterizing `User`→`LCA` *information flows* (Section 2). Our contextual privacy notion requires that user prompts include only information that is contextually appropriate, relevant, and necessary to achieve the user's intended goals when interacting with LCAs, going beyond approaches that simply protect sensitive information (Dou et al., 2023; Siyan et al., 2024). For instance, when a user is querying an LCA of a bank to locate tax forms, sharing SSN would adhere to contextual privacy, as it may be necessary for the task. On the other hand, if a user seeks advice on managing personal finances, sharing the names of family members would violate contextual privacy.

Proposed Framework: We design a framework that can protect users during their interactions with LCAs. By analyzing user inputs, detecting potentially sensitive irrelevant content, and guiding users to reformulate prompts based on contextual relevance, our framework 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.

Our main contributions include:

- We formulate the definition of contextual privacy for the specific case of `User`→`LCA` information flows, where users act as senders and LCAs as untrusted receivers;
- We apply our contextual privacy definition to analyze real-word conversation from ShareGPT (Chiang et al., 2023) and demonstrate how users unintentionally violate contextual privacy in interactions with LCAs;
- We develop a privacy safeguarding framework that acts as an intermediary between the user and LCA, and helps users identify and reformulate out-of-context information in their prompts while maintaining their intended goals;
- We design novel metrics to measure the contextual privacy and utility performance of our framework;
- We show that our privacy safeguarding framework can be implemented using a small LLM that can be locally deployed at the user side. We consider three state-or-the-art models for implementation, and compare their privacy and utility performances. Our experiments shows that lightweight models can effectively implement this framework, achieving both strong privacy protection and utility through different approaches to classify information relevant to the intended goals.

We fully contextualize our contributions with regards to existing literature in Appendix A.

2 Threat Models and Privacy Definition

Threat Model. We consider a scenario where users interact with large, remote, and untrusted LCAs through APIs. These 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.

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 Mark, 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 Gina and Emma 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	Wei'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 Sandy. Include our 13 week old daughter named Hailey and add in some humor.

We focus on a threat model where users unintentionally compromise their privacy by oversharing information. Our approach targets out-of-context *self-disclosure* by guiding users to share only contextually necessary information. By identifying unnecessary or sensitive disclosures in real time, we assist users in controlling the information they reveal, thereby reducing the risk of unintentional privacy breaches. Our approach indirectly mitigates the threat of *malicious users*, who seek to extract sensitive information from the agents by manipulating their interactions, by minimizing the amount of sensitive information exchanged during interactions.

Contextual Privacy in Conversational Agents We define the notion of *contextual privacy* in conversational agents, inspired by the Contextual Integrity (CI) theory. CI models privacy as information flow defined by the five parameters sender (who is sharing the data), subject (who the information is about), receiver (who is getting the data), context (what sort of information is being shared), and transmission principle (the conditions under which information flow is conducted) (Nissenbaum, 2004). CI evaluates whether the information flow adheres to appropriate standards governed by norms, which vary based on the specific circumstances of the interaction. Establishing privacy norms and privacy principles of CI is complex and indeed an open problem in the literature since norms are governed by societal contexts and can evolve in response to societal developments (Malkin, 2023).

Instead, we draw inspiration from the CI theory to formalize the notion of contextual privacy, focusing on the user-LCA interaction. We begin with characterizing the *information flow* between a user and an LCA by drawing on the five essential CI parameters in Table 2. We simplify the transmission principle based on the privacy directive *share information that is essential to get the answer*, similar to (Bagdasaryan et al., 2024).

After we characterize the subject and the *context* (which captures the user's intent and the key task) from the user's query along with the prior conversation history, we determine two types of sensitive attributes in the query: (a) details that are essential to answer the query, and (b) sensitive details that are not essential for answering the query. We say that a user query is *contextually private* if it does not contain any nonessential sensitive attributes. An example of essential and non-essential attributes for a query is shown in Figure 1.

3 A Framework for Safeguarding Contextual Privacy

Our goal is to develop a framework that acts as an intermediary between the user and LCA, and enables the user to detect whether their prompt incurs any contextual privacy violations, and judiciously reformulate the prompt to ensure contextual privacy. We first conduct a formative design study to guide our framework design.

User Study to Guide Our Framework Design: We conducted a *Wizard-of-Oz* formative user study to explore users' expectation of privacy when interacting with LCAs and to gather technical requirements for our framework. Following established practices in early-stage interface design research (Nielsen, 2000; Budiu, 2021; Nielsen and Landauer,

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

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.

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 B.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 framework, including distinguishing between essential and non-essential sensitive information, real-time feedback, user control over reformulations, and 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 B.

Proposed Framework: We propose a framework that acts as an intermediary between the user and the 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 C).

Once the context and subject are identified, our framework moves on to detecting sensitive information in the prompt. The framework categorizes the sensitive information into two spaces: (a) **essential information space**: sensitive details necessary to answer the user's query, (b) **non-essential information space**: sensitive details that are unnecessary for answering the query and should be kept private.

In the example of Figure 1, 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 non-essential.

Once contextually essential and non-essential information is identified, our framework improves contextual privacy by **reformulating** the prompt. This process includes removing, rephrasing, or redacting details within the non-essential 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 LCA. 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 non-essential 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.

4 Implementation and Evaluation

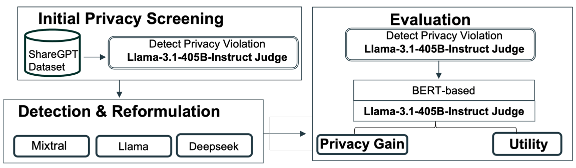


Figure 2: Experimental pipeline showing initial privacy screening, reformulation by three local models, and evaluation stages.

4.1 Contextual Privacy Evaluation of Real-World Queries

Before implementing and evaluating our framework, we first perform initial privacy analysis by evaluating an open-source version of the ShareGPT dataset (Chiang et al., 2023) to understand the prevalence of contextual privacy violations. To instantiate our formal privacy definition, we used Llama-3.1-405B-Instruct (Grattafiori et al., 2024) as judge, with a prompt designed to identify violations of contextual integrity (Appendix D.1). From over 90,000 conversations, we retain 11,305 single-turn conversations within a reasonable length range (25-2,500 words). For each conversation, the judge model assessed the context, sensitive information, and their necessity for task completion. This analysis identified approximately 8,000 conversations containing potential contextual integrity violations. To manage inference costs, we focused on cases where the judge model could successfully identify a primary context and classify essential and non-essential information attributes, yielding 2,849 conversations (25.2%) with definitive contextual privacy violations. Examples of these violations are shown in Table 1. Manual inspection of the judge's results for consistency and correctness demonstrated good classification performance with few false positives and negatives.

4.2 Implementation Details

Models. We implement our framework using a model that is significantly smaller than typical chat agents like ChatGPT, enabling users to deploy the model locally via Ollama[1] without relying on external APIs. In our experiments, we evaluate three models with different characteristics: Mixtral-8x7B-Instruct-v0.1[2] (Jiang et al., 2024), Llama-3.1-8B-Instruct[3] (Grattafiori et al., 2024), and DeepSeek-R1-Distill-Llama-8B[4] (focused on reasoning) (DeepSeek-AI et al., 2025). We refer to these models as Mixtral, Llama and Deepseek in short going forward. The local deployment of models ensures no further privacy leakage due to the framework. Although our evaluation focuses on three LLMs, our approach is model-agnostic and can be applied to other architectures. For assessment of privacy and utility, we use Llama-3.1-405B-Instruct (Grattafiori et al., 2024) as an impartial judge, which was hosted in a secure cloud infrastructure.

Experiment Setup. As discribed in the previous section, our framework processes user prompts in three stages: (a) context identification, (b) sensitive information classification, and (c) reformulation. The locally deployed model first determines the context of the conversation, identifying its domain and task (Appendix C) using the prompts in Appendix Appendix D.2 and Appendix D.3 respectively. It then detects sensitive information, categorizing it as either *essential* (required for task completion) or *non-essential* (privacy-sensitive and removable). Finally, if non-essential sensitive information is present, the model reformulates the prompt to improve privacy while preserving intent.

We implement two approaches for sensitive information classification: **dynamic classification** and **structured classification**, each reflecting different ways to operationalize our privacy frame-

[1] https://github.com/ollama/ollama

[2] https://ollama.com/library/mixtral:8x7b-instruct-v0.1-q4_0

[3] https://ollama.com/library/llama3.1:8b-instruct-fp16

[4] https://ollama.com/library/deepseek-r1:8b-llama-distill-q4_K_M

work. In the **dynamic classification approach** (see prompt used in Appendix D.4), the model determines which details are essential based on how they are used within the specific conversation. For instance, in the prompt *"I'm Jane, a single parent of two, and was just diagnosed with diabetes. I'm looking for affordable treatment options"*, the model would identify the phrases= *["diabetes"]* as the essential attributes, while *["Jane", "single parent of two","affordable"]* would be classified as non-essential. This adaptive method aligns with contextual privacy formulation, ensuring that only task-relevant details are retained. In contrast, the **structured classification approach** (see prompt used in Appendix D.5), allows to specify a predefined list of sensitive attributes (e.g., age, SSN, physical health, allergies) that should always be considered non-essential (protected), ensuring consistent enforcement of privacy policies. For the same example, this approach would flag *["physical health"]* as the essential attribute while labeling *["name", "family status", "financial condition"]* as non-essential attributes, recommending them for removal based on user-defined privacy preferences. This provides greater control over what information is considered sensitive, allowing customization while maintaining a standardized privacy framework. The predefined attribute categories follow those defined in Bagdasaryan et al. (2024).

If non-essential sensitive details are detected, the model reformulates the prompt by either removing or rewording them to minimize privacy risks while maintaining usability (see Prompt used in Appendix D.6). By evaluating both dynamic and structured classification, we demonstrate the flexibility of our framework in balancing adaptability with user-defined privacy controls.

4.3 Evaluation and Results

We evaluate our framework by measuring two key metrics: **privacy gain** and **utility**. Privacy gain quantifies how effectively sensitive information is removed during reformulation, while utility measures how well the reformulated prompt maintains the original prompt's intent. We compute these metrics using two complementary methods: an automated BERTScore-based comparison of sensitive attributes, and an LLM-based assessment that aggregates multiple evaluation aspects.

Table 3: BERT-based Evaluation of Privacy and Utility

Model	Privacy Gain ↑	Utility(BERTScore)↑
Dynamic Attribute Classification		
Deepseek	0.853	0.570
Llama	0.886	0.567
Mixtral	0.873	0.570
Structured Attribute Classification		
Deepseek	0.836	0.511
Llama	0.873	0.606
Mixtral	0.824	0.576

4.3.1 Evaluation via Attribute-based Metrics

Metrics. We measure privacy gain by computing semantic similarity between non-essential attributes between original and reformulated prompts, where similarity is computed using BERTScore (Zhang et al., 2020). Specifically, we first run the judge model on reformulated prompts to obtain non-essential sensitive attributes $\mathcal{P}^{\text{reform}}_{\text{non-ess}}$, using a prompt designed to identify contextual privacy violations (Appendix D.1). We have non-essential sensitive attributes for original prompts $\mathcal{P}^{\text{orig}}_{\text{non-ess}}$ from Section 4.1. Given sets of strings $\mathcal{P}^{\text{orig}}_{\text{non-ess}}$ and $\mathcal{P}^{\text{reform}}_{\text{non-ess}}$, privacy gain is computed as $1 - \text{BERTScore}(\mathcal{P}^{\text{orig}}_{\text{non-ess}}, \mathcal{P}^{\text{reform}}_{\text{non-ess}})$, with a score of 1.0 assigned when either set is empty. A higher privacy gain indicates better removal of sensitive information. For utility, we measure semantic similarity between essential attributes using $\text{BERTScore}(\mathcal{P}^{\text{orig}}_{\text{ess}}, \mathcal{P}^{\text{reform}}_{\text{ess}})$, where a score closer to 1.0 indicates better preservation of task-critical information. Since BERTScore works on text pairs, we match each original attribute to its closest reformulated one and compute utility as the fraction of matched attributes above a similarity threshold of 0.5.

Results. Table 3 shows that under dynamic classification, all three models achieve strong privacy scores (0.85-0.88) with comparable utility ($\sim$ 0.57), suggesting that the ability to identify context-specific sensitive information is robust across different model architectures.

The structured classification approach shows greater variation between models. While Llama achieves high scores in both privacy (0.873) and utility (0.606), structured classification generally yields slightly lower privacy scores but more variable utility. This suggests a natural trade-off: predefined categories might miss some context-specific sensitive information, yet operating within these fixed boundaries can help preserve task-relevant content. Interestingly, the similar performance patterns across different model architectures suggest that the choice between instruction-tuned and reasoning-focused approaches may be less crucial for privacy-preserving reformulation.

The success of both dynamic and structured approaches offers implementation flexibility - users can choose predefined privacy rules or context-specific protection based on their requirements. This choice, rather than model architecture, appears to be the key decision factor in deployment.

4.3.2 LLM-as-a-Judge Assessment

Setup. We use Llama-3.1-405B-Instruct as a judge to provide a complementary evaluation of privacy and utility across 100 randomly selected queries per model (6×100 total). Given the high computational cost of LLM-based inference, this targeted sampling allows us to validate key trends observed in the attribute-based evaluation while minimizing overhead. Privacy gain is computed by asking the judge to evaluate privacy leakage, coverage, and retention, while utility is computed by measuring query relevance, response validity, and cross-relevance. These binary evaluations are averaged to produce final privacy gains and utility scores. See Appendix D.7 for detailed prompts and evaluation criteria.

Results. The LLM-based assessment shows generally higher utility scores (0.82-0.86) across all models compared to BERTScore-based evaluation, while maintaining similar privacy levels (0.80-0.86). This difference can be attributed to how attributes are detected and compared—BERTScore evaluates exact semantic matches between attributes, while the LLM judge takes a more holistic view of information preservation. For instance, when essential information is restructured (e.g., "my friend Mark" split into separate attributes), BERTScore may indicate lower utility despite semantic equivalence.

The LLM evaluation confirms the effectiveness of both classification approaches, with dynamic classification showing slightly more consistent performance across models. Llama maintains its strong performance under both approaches (privacy gain: ~ 0.85, utility score: ~ 0.86), reinforcing its reliability for privacy-preserving reformulation.

Table 4: LLM-as-a-Judge Evaluation of Privacy and Utility

Model	Privacy Gain ↑	Utility Score↑
Dynamic Attribute Classification		
Deepseek	0.802	0.845
Llama	0.858	0.861
Mixtral	0.848	0.838
Structured Attribute Classification		
Deepseek	0.815	0.825
Llama	0.855	0.858
Mixtral	0.845	0.828

4.3.3 Example Reformulations and Trade-offs

Setup. Table 5 presents a set of diverse example reformulations illustrating our framework's ability to balance privacy and utility across different scenarios. These examples highlight both ideal cases—where reformulation effectively preserves both privacy and utility—and more challenging ones where trade-offs are unavoidable.

Results. Our framework successfully removes personal identifiers while preserving task relevance, as seen in the third example (privacy gain = 0.5, utility score = 0.83). In creative requests like the Valentine's poem (second example), removing personal details reduces privacy risks but slightly impacts personalization utility = 0.5).

Some contexts resist reformulation. The last conversation (privacy gain = 0.0, utility score = 0.0) highlights cases where the entire prompt is inherently sensitive, requiring alternative privacy measures beyond text transformation. This is not a failure of our approach but an indicator of when reformulation alone is insufficient.

These examples reinforce that privacy-preserving reformulation is a trade-off, not a one-size-fits-all solution.

5 Discussion and Conclusion

Drawing ideas from the contextual integrity theory, we defined the notion of contextual privacy for users interacting with LLM-based conversation agents. We proposed a framework, grounded in our contextual privacy formulation, that acts as an intermediary between the user and the agent, and carefully reformulates user prompts to preserve contextual privacy while preserving the utility.

This work serves as an initial step in exploring privacy protection in user interactions with conversational agents. There are several directions that

Table 5: Examples of original and reformulated prompts with privacy and utility scores. These examples illustrate how our framework removes sensitive details while preserving task intent. The high privacy, high utility cases demonstrate successful reformulation, while the low privacy, low utility example highlights cases where effective reformulation is difficult or unnecessary.

Original Prompt	Reformulated Prompt	Privacy Gain	Utility (BERTScore)
My friend Mark 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?	Someone is looking for a job where they can use ML and Python skills. Do you have any advice?	**1.0**	**1.0**
Please write a poem for my wife for Valentine's Day. Her name is Gili. Please use rhymes. Please make it sarcastic, but not too much so. A bit funny but romantic.	Write a poem for Gili, incorporating sarcasm at an appropriate level and maintaining a tone that is both funny and romantic	**1.0**	**0.5**
I am writing a character letter to a judge in hopes of having charges against my friend Amber to be dropped. Write a letter in a positive tone to a judge speaking to Amber's good character ...... How you know me, my ties to the community my profession and stating a dismissal is imperative to my earning potential ... I know amber from "Teach For America". I know her to be a passionate educator who cares about uplifting the youth and her community. She should have the charges dropped so that she can increase her community engagement and earning potential for her family.	Write a character letter in a positive tone to a judge speaking to someone's good character, highlighting their community ties and profession, and stating that a dismissal is imperative to their earning potential. I know this person from a professional program and they have been a passionate advocate for the betterment of their community. The charges should be dropped so that they can increase their engagement in the community and improve their financial situation.	**0.5**	**0.83**
Sunny Balwani : I worked for 6 years day and night to help you. Elizabeth Holmes : I was just thinking about texting you in that minute by the way	Sunny Balwani : I am responsible for everything at Theranos. Elizabeth Holmes :	**0.0**	**0.0**

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 and can be sensitive to the prompt. 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.

Limitations

Contextual integrity is a relatively new and fluid notion of privacy. Ours is also one of the very early works exploring this space from the standpoint of LLM-based conversational agents. Naturally, this leads to a number of challenges, some of which are beyond the scope of the work and should be addressed in the future. Like we discussed before, establishing privacy norms and principles in CI itself is complex and dependent on societal contexts, which is why we restrict ourselves to a practical and useful variation of the idea. However, developing templates for implementing CI under various societal contexts deserves significant attention from the research community in the future.

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.

References

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

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

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

Raluca Budiu. 2021. Why 5 participants are okay in a qualitative study, but not in a quantitative one. Accessed: November 27, 2024.

Alycia N Carey, Karuna Bhaila, Kennedy Edemacu, and Xintao Wu. 2024. Dp-tabicl: In-context learning with differentially private tabular data. *arXiv preprint arXiv:2403.05681*.

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

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

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. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

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

Lynn Chua, Badih Ghazi, Yangsibo Huang, Pritish Kamath, Daogao Liu, Pasin Manurangsi, Amer Sinha, and Chiyuan Zhang. 2024. Mind the privacy unit! user-level differential privacy for language model fine-tuning. *arXiv preprint arXiv:2406.14322*.

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

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

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

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

Kennedy Edemacu and Xintao Wu. 2024. Privacy preserving prompt engineering: A survey. *arXiv preprint arXiv:2404.06001*.

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

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim

Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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

Junyuan Hong, Jiachen T Wang, Chenhui Zhang, Zhangheng Li, Bo Li, and Zhangyang Wang. 2023. Dp-opt: Make large language model your privacy-preserving prompt engineer. *arXiv preprint arXiv:2312.03724*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

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

Abhishek Kumar, Tristan Braud, Young D Kwon, and Pan Hui. 2020. 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.

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

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

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. 2021. Large language models can be strong differentially private learners. *arXiv preprint arXiv:2110.05679*.

Pierre Lison, Ildikó Pilán, David Sánchez, Montserrat Batet, and Lilja Øvrelid. 2021. 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)*, pages 4188–4203.

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

Nathan Malkin. 2023. Contextual Integrity, Explained: A More Usable Privacy Definition. *IEEE Security & Privacy*, 21(01):58–65.

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

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

R Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. 2023. 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.

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

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

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

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

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

Jakob Nielsen. 2000. Why you only need to test with 5 users. Accessed: November 27, 2024.

Jakob Nielsen and Thomas K Landauer. 1993. 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*, pages 206–213.

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

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

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

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

Ashok Kumar Reddy Sadhu, Maksym Parfenov, Denis Saripov, Maksim Muravev, and Amith Kumar Reddy Sadhu. 2024. 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.

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

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

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

Xinyu Tang, Richard Shin, Huseyin A Inan, Andre Manoel, Fatemehsadat Mireshghallah, Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, and Robert Sim. 2023. Privacy-preserving in-context learning with differentially private few-shot generation. *arXiv preprint arXiv:2309.11765*.

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

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

Tong Wu, Ashwinee Panda, Jiachen T Wang, and Prateek Mittal. 2023. Privacy-preserving in-context learning for large language models. *arXiv preprint arXiv:2305.01639*.

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

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

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. 2021. Differentially private fine-tuning of language models. *arXiv preprint arXiv:2110.06500*.

Xiang Yue, Huseyin A Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. 2022. Synthetic text generation with differential privacy: A simple and practical recipe. *arXiv preprint arXiv:2210.14348*.

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

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

Liang Zhang, Bingcong Li, Kiran Koshy Thekumparampil, Sewoong Oh, and Niao He. 2024a. Dpzero: Private fine-tuning of language models without backpropagation. In *Forty-first International Conference on Machine Learning*.

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

Zhiping Zhang, Michelle Jia, Hao-Ping Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. 2024b. "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*, pages 1–26.

A Related Work

We fully contextualize our contributions in regard to existing literature here.

LLM Privacy-Preserving Techniques. A significant body of research on privacy preservation in LLMs has focused on the training phase (Zhang et al., 2024a; Chua et al., 2024; Yu et al., 2021; Yue et al., 2022; Li et al., 2021). Techniques like differential privacy (DP) (Dwork et al., 2006) have been used to prevent LLMs from memorizing sensitive information during training. Additionally, data sanitization strategies, such as deduplication and anonymization, have been used to reduce privacy risks by removing sensitive data from training data (Lison et al., 2021; Kandpal et al., 2022). After training, machine unlearning methods have emerged to help eliminate any retained private data (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 received less attention, with limited approaches, such as PII detection and DP decoding, targeting the risks of exposing sensitive information in real-time interactions with LLMs (Majmudar et al., 2022; Carey et al., 2024; Wu et al., 2023; Tang et al., 2023; Hong et al., 2023; Edemacu and Wu, 2024). Recently, Mireshghallah et al. (2023) highlighted this gap, showing that LLMs often fail to protect private information in context and emphasizing the need for better privacy-preserving techniques. Our approach addresses this need by offering real-time, context-aware privacy guidance during user interactions, allowing individuals to better manage what information they disclose during conversations with LLMs.

Privacy Risks in Human-LLM Interactions. Self-disclosure during human-machine interactions can result in unintended sharing of sensitive information. For example, Ravichander and Black (2018) found that users tend to reciprocate with automated systems, revealing more personal information over time. Building on this, Zhang et al. (2024b) examined the privacy risks faced by users interacting with LLMs, showing that human-like responses can encourage sensitive disclosures, complicating privacy management. Mireshghallah et al. (2024) further advanced this discussion by highlighting the limitations of PII detection systems, showing that users often disclose sensitive information that goes beyond PII (Cummings et al., 2023; Dou et al., 2023). Our work builds on these efforts by showing that users frequently disclose unnecessary information during interactions with LLMs, which can be contextually sensitive and unrelated to their intended goals. We develop a system that detects such information and offers reformulation suggestions to guide users toward more privacy-aware interactions.

Data Minimization in ML. The principle of data minimization, central to privacy regulations like GDPR (Voigt and Von dem Bussche, 2017), has recently been a key focus in ML research. For example, Ganesh et al. (2024) formalized data minimization within an optimization framework for reducing data collection while maintaining model performance. Tran and Fioretto (2024) expanded on this by showing that individuals can disclose only a small subset of their features without compromising accuracy, thus minimizing the risk of data leakage. While both approaches focus on reducing the amount of data processed at inference time, our work applies data minimization in real time, guiding users to share only necessary information with LLMs. We integrate contextual integrity to ensure that the disclosed information aligns with the context of the conversation, ensuring GDPR compliance through a user-driven, context-aware approach.

Operationalizing Contextual Integrity (CI). Research on contextual privacy in LLMs is rapidly expanding. For instance, Mireshghallah et al. (2023) introduced a benchmark to evaluate the privacy reasoning abilities of LLMs at varying levels of complexity, while Shvartzshnaider et al. (2024) proposed a comprehensive framework using CI to assess privacy norms encoded in LLMs across different models and datasets. CI has also been integrated into various practical systems to safeguard privacy across diverse domains. For example, Shvartzshnaider et al. (2019) employed CI to detect privacy leaks in email communications, and Kumar et al. (2020) applied CI to provide mobile users with real-time privacy risk alerts. In smart home ecosystems, Malkin et al. (2022); Abdi et al. (2021) used CI to analyze and enforce privacy norms. Hartmann et al. (2024) considered scenarios where a local model queries a larger remote model, leveraging CI to ensure only task-relevant data is shared. Similarly, Bagdasaryan et al. (2024) used CI to restrict AI assistants' access to only the information necessary for a given task, and Ghalebikesabi et al. (2024) applied CI to ensure form-filling assistants follow contextual privacy norms when sharing user information. While these studies focus on aligning AI assistants' actions with privacy norms, our work shifts the perspective 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 manage privacy risks. This user-centered approach not only protects sensitive information during AI interactions but also promotes long-term privacy awareness—an aspect often overlooked in system-oriented solutions.

B User Study to Guide System Design

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

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 B.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.

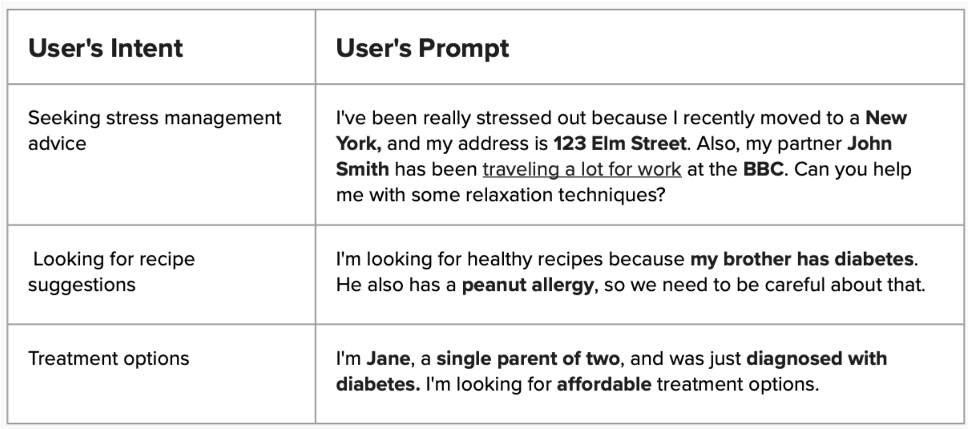
- **User control and real-time feedback**. Real-time feedback and user control over editing flagged prompts were highly valued. Participants preferred having the system automatically generate reformulations, but they wanted the ability to make any necessary final adjustments. **Design impact**: We implemented 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.

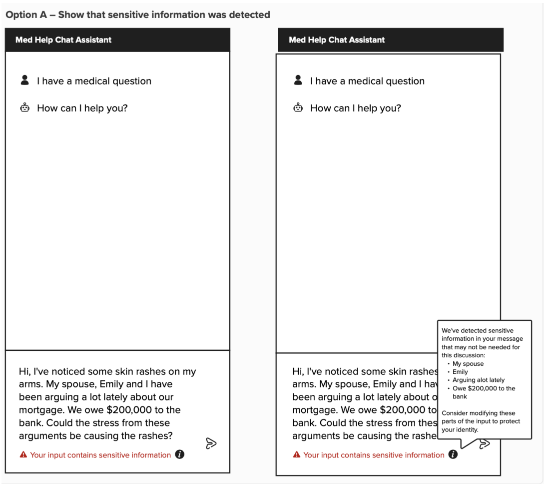
- **Broader application**. A few participants suggested applying the tool to other contexts beyond LLM chat interfaces, such as search engines. **Design impact**: This feedback highlights the importance of managing sensitive information and the broader applicability of our approach to other contexts.

B.1 User Study Mockups

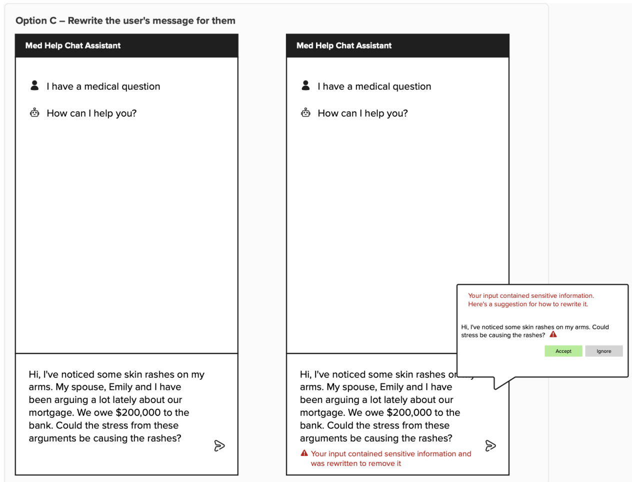


User's Intent	User's Prompt
Seeking stress management advice	I've been really stressed out because I recently moved to a **New York,** and my address is **123 Elm Street**. Also, my partner **John Smith** has been traveling a lot for work at the **BBC**. Can you help me with some relaxation techniques?
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.

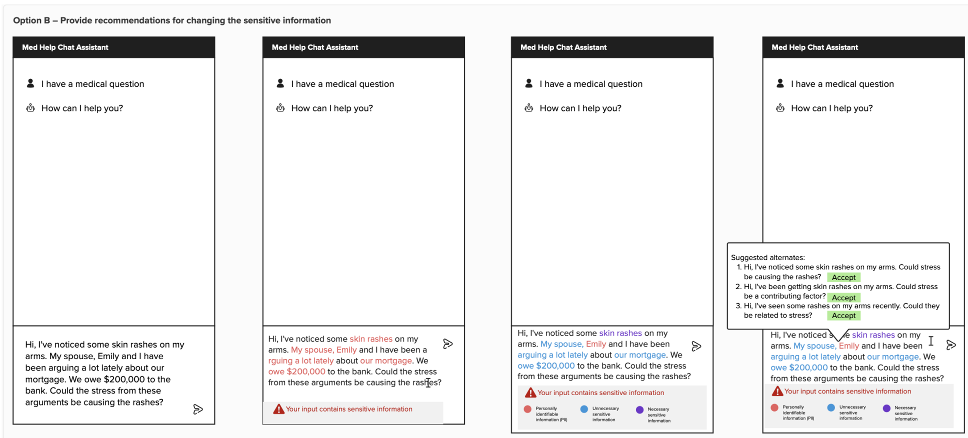
(a) Examples of unintentional disclosures shown to participants



(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

C Domains and Tasks

Table 6 shows the list of Domain and Tasks Categories for Intent Detection.

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

D Prompts

D.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:

Detection of Contextual Privacy Violations

Task: Identify and classify sensitive information in the given text under three categories:

- **"primary context"** - The key task, question, or request present in the text.
- **"attributes essential to the context"** - Necessary details directly relevant to the primary context.
- **"sensitive attributes not essential to the context"** - Other sensitive information unrelated to the primary context.

Examples:
Example 1: Text: *"Hi, my name is David 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": [
"my name is David Johns",
"stressed about my financial situation",
"in debt",
"struggling to pay my medical bills"
]}
```
Example 2: Text: *"I recently moved to San Francisco and I'm having trouble finding a job. My email is john.doe@gmail.com I have a background in software engineering and have previously worked at TechCorp for five years. Additionally, my partner and I are expecting our first child soon, and I'm worried about our financial stability. Give me some tips for finding a job."*
Output:
```
{"primary context": ["give some tips for finding a job"],
"attributes essential to the context": [
"having trouble finding a job",
"background in software engineering",
"worked at TechCorp for five years",
"give tips for finding a job"
],
"sensitive attributes not essential to the context": [
"recently moved to San Francisco",
"my email is john.doe@gmail.com",
"expecting our first child",
"worried about our financial stability"
]}
```
Now classify the text below:
Text: {input_text}

D.2 Prompt Template for Intent Detection

Intent Detection Prompt Template

- if .System You are an intent detection assistant that identifies the primary intent or domain of a user's query. You must classify the intent into exactly one of the available categories below. end

User Prompt: Identify the **PRIMARY** intent/domain of the text following these rules:

1. Choose the best category from the provided list.
2. If no exact match exists, select the closest category.
3. DO NOT create new categories or provide explanations—return only a category name.

Available Categories:

- **Health_And_Wellness** - Health conditions, treatments, therapy, fitness, nutrition.
- **Financial_And_Corporate** - Banking, budgeting, loans, insurance, markets.
- **Employment_And_Applications** - Job searches, workplace, professional matters.
- **Academic_And_Education** - School, grades, learning resources, teaching.
- **Legal** - Cases, contracts, laws, government benefits.
- **Personal_Relationships** - Family, friends, social life, childcare.
- **Travel** - Trip planning, transport, documents, accommodation.
- **Fandom** - Books, movies, TV, games, fan communities.
- **Hobbies_And_Habits** - Leisure, gaming, sports, collecting.
- **Sexual_And_Erotic** - Preferences, relationships, health, adult topics.
- **Politics** - Policies, elections, ideologies, civic matters.
- **Religion** - Beliefs, spirituality, religious events, worship.

Examples:

- **Text:** "Looking for a new job after being laid off from my tech company" → **Employment_And_Applications**
- **Text:** "Need advice for managing my anxiety and depression" → **Health_And_Wellness**
- **Text:** "Want to discuss the latest episode of Game of Thrones" → **Fandom**
- **Text:** "Planning a family vacation to Europe next summer" → **Travel**

Now classify this text. Return only a category name: Text: {input_text}

1344
D.3 Prompt Template for Task Detection

Task Detection Prompt Template

Task: Identify the primary task or action type from the text using the predefined categories.
Rules:

- Select the most appropriate category from the list.
- If no exact match exists, choose the closest category.
- Return only the category name—no explanations or additional text.

Available Categories:

- Summarization
- Prompt_Generation_For_AI_Models
- Story_And_Script_Generation
- Song_And_Poem_Generation
- Character_Description_Generation
- Code_Generation
- Code_Editing_And_Debugging
- Communication_Generation
- Non_Fictional_Document_Generation
- Text_Editing
- Comparison_Ranking_And_Recommendation
- Brainstorming_And_Idea_Generation
- Information_Retrieval
- Problem_Solving
- Explanation_And_Practical_Advice
- Personal_Advice
- Back_And_Forth_Role_Playing
- Answering_Multiple_Choice_Questions
- Translation
- General_Chitchat

Examples:
Example 1: Text: *"Can you help me write a prompt that will make GPT-4 better at generating poetry?"* **Task:** `Prompt_Generation_For_AI_Models`
Example 2: Text: *"I need help debugging this Python function that keeps giving me a TypeError."* **Task:** `Code_Editing_And_Debugging`
Example 3: Text: *"Write me a story about a detective solving a mystery in space."* **Task:** `Story_And_Script_Generation`
Example 4: Text: *"How do I tell my roommate that I'm moving out next month?"* **Task:** `Personal_Advice`
Now classify the text below:
1345 **Text:** {input_text}

D.4 Dynamic Prompt Template for Sensitive Information Detection

Dynamic Sensitive Information Detection Prompt Template

ESSENTIAL PROMPT:
- if .System You are a privacy-focused assistant. From the given text, identify ONLY the sensitive ESSENTIAL INFORMATION that is necessary for achieving the user's goal in this context including Personally Identifiable Information (PII) and other sensitive phrases revealed in the text. end

User Prompt: Your task is to identify ONLY the essential sensitive information that is necessary for achieving the user's goal in this context of {intent}.
Criteria:

- Sensitive information that falls specifically in the context of {intent}.
- Essential for getting a helpful response.
- Cannot be removed without significantly impacting the ability to help the user.

Return format: ESSENTIAL INFORMATION: [item 1, item 2, item 3]

Examples:
Example 1: Context: Job Search for Employment_And_Applications
Input Text: "My friend Justin, who was laid off from Google, is looking for a job where he can use ML and Python. Do you have any advice for him?"
ESSENTIAL INFORMATION: `["looking for job using ML and Python skills"]`
Example 2: Context: Medical Advice for Health_And_Wellness
Input Text: "I need advice managing my type 2 diabetes. My doctor prescribed Metformin 500mg twice daily last week. My sister also has diabetes and suggested trying a keto diet."
ESSENTIAL INFORMATION: `["type 2 diabetes", "prescribed Metformin 500mg twice daily"]`

Now analyze the following text and extract ONLY the ESSENTIAL INFORMATION: Context: {intent} **Input Text:** {text}

NON-ESSENTIAL PROMPT:
- if .System You are a privacy-focused assistant. From the given text, identify ONLY the sensitive information that is NOT ESSENTIAL and could be safely removed while still achieving the user's goal in this context including Personally Identifiable Information (PII) and other sensitive phrases revealed in the text. end

User Prompt: Your task is to identify information that is NOT ESSENTIAL for the user's goal in this context of {intent}.
Criteria:

- All sensitive information that does not directly pertain to {intent}.
- Not necessary for achieving the intended goal.
- Could be removed or generalized while still getting a helpful response.

Return format: NON-ESSENTIAL INFORMATION: [item 1, item 2, item 3]

Examples:
Example 1:
Context: Job Search for Employment_And_Applications
Input Text: "My friend Justin, who was laid off from Google, is looking for a job where he can use ML and Python. Do you have any advice for him?"
NON-ESSENTIAL INFORMATION: `["my friend", "Justin", "laid off from Google"]`
Example 2:
Context: Medical Advice for Health_And_Wellness
Input Text: "I need advice managing my type 2 diabetes. My doctor prescribed Metformin 500mg twice daily last week. My sister also has diabetes and suggested trying a keto diet."
NON-ESSENTIAL INFORMATION: `["prescribed last week", "sister has diabetes", "suggested trying a keto diet"]`

Now analyze the following text and extract ONLY the NON-ESSENTIAL INFORMATION:
Context: {intent} **Input Text:** {text}

D.5 Structured Prompt Template for Sensitive Information Detection

Structured Sensitive Information Detection Prompt Template

ESSENTIAL PROMPT:
- if .System You are a privacy-focused assistant. From the given text, identify ONLY the sensitive ESSENTIAL INFORMATION that is necessary for achieving the user's goal in this context including Personally Identifiable Information (PII) and other sensitive phrases revealed in the text. end
User Prompt: Your task is to identify ONLY the essential sensitive information that is necessary for achieving the user's goal in this context of {intent}.
Use ONLY these categories: `[age, driver license, phone number, SSN, allergies, exercise hours, medications, mental health, physical health, disabilities, family history, diet type, favorite food, favorite hobbies, pet ownership, movie prefs, relationship status, religious beliefs, sexual orientation, vacation prefs, name, email, address, ethnicity, gender, smoker, financial situation, legal, employment, dates]`
Criteria:

- Sensitive information that falls specifically in the context of {intent}.
- Essential for getting a helpful response.
- Cannot be removed without significantly impacting the ability to help the user.

Return format: ESSENTIAL INFORMATION: [item 1, item 2, item 3]

Examples:
Example 1:
Context: Employment_And_Applications
Input Text: "My friend Justin, who was laid off from Google, is looking for a job where he can use ML and Python. Do you have any advice for him?"
ESSENTIAL INFORMATION: `["employment"]`
Example 2: Context: Medical_And_Health
Input Text: "I need advice managing my type 2 diabetes. My doctor prescribed Metformin 500mg twice daily last week. My sister also has diabetes and suggested trying a keto diet."
ESSENTIAL INFORMATION: `["physical health", "medications", "diet type"]`

Now identify the essential attributes from the predefined list:
Context: {intent} **Input Text:** {text}

NON-ESSENTIAL PROMPT:
- if .System You are a privacy-focused assistant. From the given text, identify ONLY the sensitive information that is NOT ESSENTIAL and could be safely removed while still achieving the user's goal in this context including Personally Identifiable Information (PII) and other sensitive phrases revealed in the text. end
User Prompt: Your task is to identify information that is NOT ESSENTIAL for the user's goal in this context of {intent}.
Use ONLY these categories: `[age, driver license, phone number, SSN, allergies, exercise hours, medications, mental health, physical health, disabilities, family history, diet type, favorite food, favorite hobbies, pet ownership, movie prefs, relationship status, religious beliefs, sexual orientation, vacation prefs, name, email, address, ethnicity, gender, smoker, financial situation, legal, employment, dates]`
Criteria:

- Present or implied in the text.
- Not necessary for achieving the user's goal.
- Could be safely removed or generalized while still getting a helpful response.

Return format: NON-ESSENTIAL INFORMATION: [item 1, item 2, item 3]

Examples:
Example 1:
Context: Employment_And_Applications
Input Text: "My friend Justin, who was laid off from Google, is looking for a job where he can use ML and Python. Do you have any advice for him?"
NON-ESSENTIAL INFORMATION: `["name", "employment"]`
Example 2:
Context: Medical_And_Health
Input Text: "I need advice managing my type 2 diabetes. My doctor prescribed Metformin 500mg twice daily last week. My sister also has diabetes and suggested trying a keto diet."
NON-ESSENTIAL INFORMATION: `["family history", "diet type"]`

Now identify the non-essential attributes from the predefined list:
Context: {intent} **Input Text:** {text}

D.6 Prompt Template For Reformulation

Privacy-Preserving Reformulation Prompt

System Prompt: You are a privacy-conscious assistant that helps reformulate text to preserve privacy while maintaining the original query structure.

User Instructions: Reformulate this text to remove sensitive or unnecessary information while preserving the original query/request structure. The reformulation should:

- Keep the same type of request (question remains question, request remains request, advice request remains advice request).
- Remove or generalize non-essential information.
- Maintain the user's position as the requester.
- **NOT** provide an answer or implementation.
- Keep all essential information.
- Maintain the original intent.
- Do **not** provide explanations, justifications, or unnecessary extra text.

Original Text: {text} **Context:** {intent} **Essential Information:** {essential_info} **Information to Remove:** {removable_info}

Examples: Original: "What's a message that I, as a cool college student, can say to someone whom I want to meet?" **Reformulated:** "What's a good way to initiate a conversation with someone I'd like to connect with?"
Original: "I need help writing an email to Professor Smith about missing class yesterday due to food poisoning." **Reformulated:** "I need help writing an email about a class absence due to illness."
Original: "Can you suggest a birthday gift for my sister Sarah who loves painting and just moved to Paris?" **Reformulated:** "Can you suggest a birthday gift for someone who enjoys painting?"
Original: "I need advice on how to tell my girlfriend Jane that I lost my job at Google and we might need to cancel our vacation to Hawaii next month." **Reformulated:** "I need advice on how to discuss a job loss and its financial impact with my partner."
Original: "Looking for advice on managing my diabetes. I take Metformin twice daily and my sister suggested trying a keto diet."
Reformulated: "Looking for advice on managing diabetes and considering dietary changes."
Original: "Can you give me advice on growing my TikTok following? I currently have 500 followers and post dance videos every day."
Reformulated: "Can you give me advice on growing a social media following?"

Now, provide only the reformulated text on a new line:

D.7 LLM-as-a-Judge Evaluation Prompt Template

LLM-as-a-Judge Evaluation Prompt

System Prompt: You are an expert evaluator of contextual privacy in large language models ensuring there are no contextual privacy violations. Given a set of original and reformulated queries and responses, along with extracted attributes, assess privacy and utility across the following dimensions:

- **Privacy Non-Leakage**: Does the reformulated query successfully remove non-essential sensitive attributes?
- **Privacy Retention**: Does the reformulated query avoid introducing new sensitive attributes?
- **Privacy Coverage**: Does the reformulated query retain all essential sensitive attributes?
- **Query Relevance**: Does the reformulated query maintain the original user intent?
- **Response Relevance**: Is the reformulated response relevant compared to the original response?
- **Cross-Relevance**: Are responses generated from the original prompt still valid for the reformulated prompt?

Example Evaluations:
Example 1: Original Query: "I'm trying to prepare a healthy meal plan for my family because my wife, Susan, is on a special diet due to her recent diagnosis of high blood pressure. We're looking for recipes that are low in sodium and heart-healthy."
Original Response: "Try incorporating recipes that focus on lean proteins, vegetables, and whole grains. A heart-healthy diet often includes dishes like grilled chicken with quinoa and steamed broccoli."
Reformulated Query: "I'm looking for healthy meal plans that are low in sodium and suitable for someone with high blood pressure."
Reformulated Response: "Focus on recipes with lean proteins, vegetables, and whole grains. Dishes like grilled chicken with quinoa and steamed broccoli are often recommended."
Evaluation Output: { "privacy non-leakage": true,
"privacy retention": true,
"privacy coverage": false,
"query relevance": true,
"response relevance": true,
"cross-relevance": true }

Now Evaluate the Following:
Original Query: {original_query} **Original Response:** {original_response} **Reformulated Query:** {reformulated_query} **Reformulated Response:** {reformulated_response}
Primary Context in Original Query: {original_primary_context} **Essential Attributes in Original Query:** {original_related_context}
Sensitive Non-Essential Attributes in Original Query: {original_not_related_context}
Primary Context in Reformulated Query: {reformulated_primary_context} **Essential Attributes in Reformulated Query:** {reformulated_related_context} **Sensitive Non-Essential Attributes in Reformulated Query:** {reformulated_not_related_context}

Return only a JSON Output with the following keys: Privacy Non-Leakage, Privacy Retention, Privacy Coverage, Query Relevance, Response Relevance, Cross-Relevance, Answerability, Making Sense.
<|Assistant|>

E User Prompts Before And After Reformulation (Table 7)

Table 7: 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.