Check for updates

CLEAR: Towards Contextual LLM-Empowered Privacy Policy Analysis and Risk Generation for Large Language Model Applications

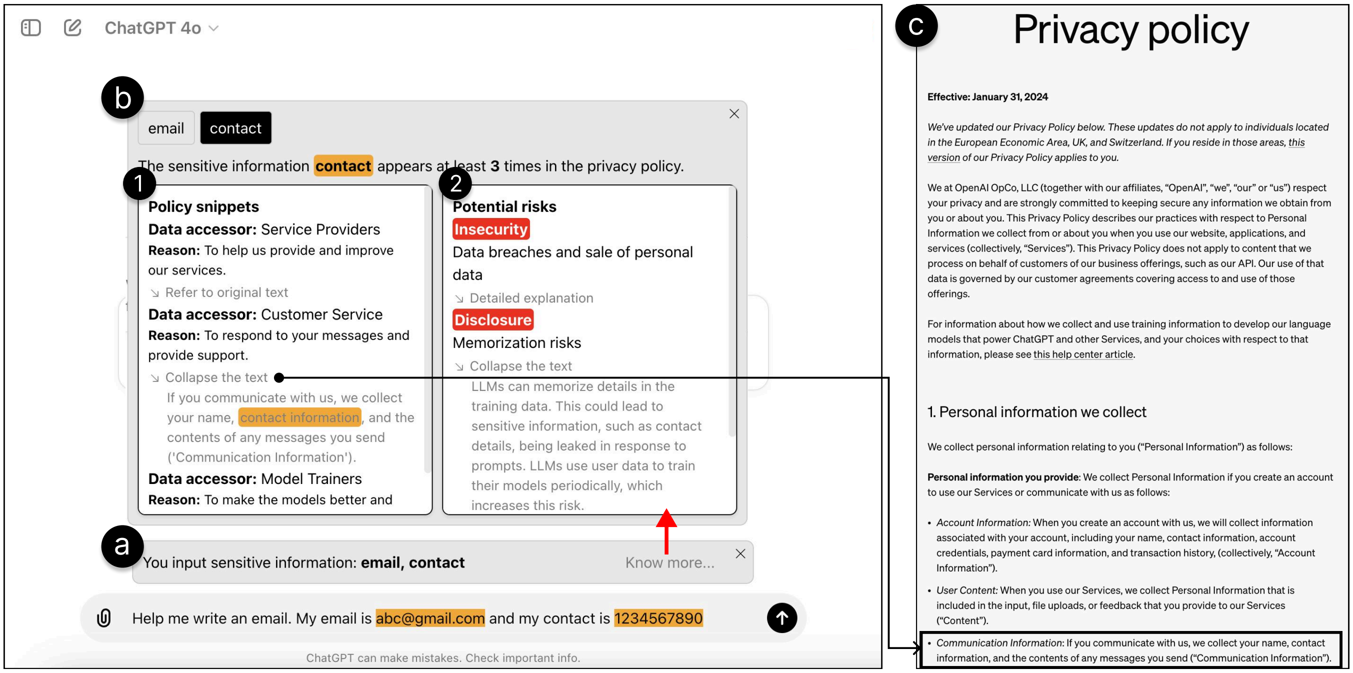
Chaoran Chen
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, Indiana, USA
cchen25@nd.edu

Daodao Zhou
Department of Computer Science
Virginia Tech
Blacksburg, Virginia, USA
dd19@vt.edu

Yanfang Ye
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, Indiana, USA
yye7@nd.edu

Toby Jia-Jun Li
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, Indiana, USA
toby.j.li@nd.edu

Yaxing Yao
Department of Computer Science
Virginia Tech
Blacksburg, Virginia, USA
yaxing@vt.edu



The sensitive information **contact** appears at least **3** times in the privacy policy.

Policy snippets

Data accessor: Service Providers
Reason: To help us provide and improve our services.
↘ Refer to original text

Data accessor: Customer Service
Reason: To respond to your messages and provide support.
↘ Collapse the text
If you communicate with us, we collect your name, contact information, and the contents of any messages you send ('Communication Information').

Data accessor: Model Trainers
Reason: To make the models better and

Potential risks

Insecurity
Data breaches and sale of personal data
↘ Detailed explanation

Disclosure
Memorization risks
↘ Collapse the text
LLMs can memorize details in the training data. This could lead to sensitive information, such as contact details, being leaked in response to prompts. LLMs use user data to train their models periodically, which increases this risk.

You input sensitive information: **email, contact**

Help me write an email. My email is abc@gmail.com and my contact is 1234567890

ChatGPT can make mistakes. Check important info.

Privacy policy

Effective: January 31, 2024

We've updated our Privacy Policy below. These updates do not apply to individuals located in the European Economic Area, UK, and Switzerland. If you reside in those areas, this version of our Privacy Policy applies to you.

We at OpenAI OpCo, LLC (together with our affiliates, "OpenAI", "we", "our" or "us") respect your privacy and are strongly committed to keeping secure any information we obtain from you or about you. This Privacy Policy describes our practices with respect to Personal Information we collect from or about you when you use our website, applications, and services (collectively, "Services"). This Privacy Policy does not apply to content that we process on behalf of customers of our business offerings, such as our API. Our use of that data is governed by our customer agreements covering access to and use of those offerings.

For information about how we collect and use training information to develop our language models that power ChatGPT and other Services, and your choices with respect to that information, please see this help center article.

1. Personal information we collect

We collect personal information relating to you ("Personal Information") as follows:

Personal information you provide: We collect Personal Information if you create an account to use our Services or communicate with us as follows:

- *Account Information:* When you create an account with us, we will collect information associated with your account, including your name, contact information, account credentials, payment card information, and transaction history, (collectively, "Account Information").
- *User Content:* When you use our Services, we collect Personal Information that is included in the input, file uploads, or feedback that you provide to our Services ("Content").
- *Communication Information*: If you communicate with us, we collect your name, contact information, and the contents of any messages you send ("Communication Information").

Figure 1: Overview of CLEAR, a contextual LLM-empowered privacy policy analysis and risk generation tool that automatically detects and displays concrete privacy policy snippets and potential risks relevant to user input sensitive information. In the above example, a user inputted the personal email and contact information in ChatGPT. CLEAR identified and displayed this information in a pop-up (a). After the user clicked "Know more...", the pop-up expanded and showed the relevant privacy policy snippets (b1) that are extracted from the lengthy privacy policy (c), as well as the potential privacy risks envisioned by LLM (b2).

Abstract

The rise of end-user applications powered by large language models (LLMs), including both conversational interfaces and add-ons to

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

IUI '25, Cagliari, Italy

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1306-4/25/03
https://doi.org/10.1145/3708359.3712156

existing graphical user interfaces (GUIs), introduces new privacy challenges. However, many users remain unaware of the risks. This paper explores methods to increase user awareness of privacy risks associated with LLMs in end-user applications. We conducted five co-design workshops to uncover user privacy concerns and their demand for contextual privacy information within LLMs. Based on these insights, we developed CLEAR (Contextual LLM-Empowered Privacy Policy Analysis and Risk Generation), a just-in-time contextual assistant designed to help users identify sensitive information, summarize relevant privacy policies, and highlight potential risks when sharing information with LLMs. We evaluated the usability and usefulness of CLEAR across two example domains: ChatGPT and the Gemini plugin in Gmail. Our findings demonstrated that CLEAR is easy to use and improves users' understanding of data practices and privacy risks. We also discussed LLM's duality in posing and mitigating privacy risks, offering design and policy implications.

CCS Concepts

• Security and privacy → Privacy protections; **• Human-centered computing → Interactive systems and tools**.

Keywords

large language model, privacy awareness, privacy intervention, privacy literacy

ACM Reference Format:
Chaoran Chen, Daodao Zhou, Yanfang Ye, Toby Jia-Jun Li, and Yaxing Yao. 2025. CLEAR: Towards Contextual LLM-Empowered Privacy Policy Analysis and Risk Generation for Large Language Model Applications. In *30th International Conference on Intelligent User Interfaces (IUI '25), March 24–27, 2025, Cagliari, Italy.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3708359.3712156

1 Introduction

As Large Language Models (LLMs) are being integrated into many end-user-facing applications of computing and fundamentally reshaping how end-users interact with computers, they introduce significant new privacy risks [60]. Past research has highlighted distinct privacy risks associated with the widespread application of LLMs. For example, LLMs are trained on extensive datasets that may include users' conversation histories, posing *memorization risks.* This risk involves the potential for LLMs to inadvertently reveal sensitive information from their training data when prompted, a problem documented in previous studies [6]). Furthermore, with their powerful reasoning capabilities, LLMs can extract personal information from seemingly benign queries (e.g., inferring users' age based on a prompt that contains lifestyle information [47]), a risk referred to as *inference risks.* There are also *disclosure risks*, where users may share more personal information than they normally would due to the LLMs' human-like interactions, fostering a false sense of trust [25, 60, 62].

In this paper, we focus exclusively on scenarios where users directly interact with LLMs, excluding those scenarios in which LLMs are used behind the scenes to analyze user data [10, 36] or substitute traditional machine learning models in back-end processes [50]. Our analysis considers two most popular user interactions with LLMs—when users directly interact with LLMs through conversational interfaces (e.g., ChatGPT) and when users interact with LLM-based add-ons in existing graphical user interface (GUI) applications (e.g., Gemini plugin in Gmail, Copilot in Microsoft Office).

Despite the emergence of these privacy risks, most users either remain unaware or misunderstand these issues [53, 60]. Users tend to trust LLMs like ChatGPT due to their human-like interactions and high utility, making them more likely to disclose personal information [60]. They often prioritize concerns about intellectual property leaks over personal privacy risks [60]. As a result, users frequently share private (e.g., emails) or semi-private (e.g., Facebook group posts) information with LLMs [30]. For instance, Zhang et al. identified numerous pieces of private data—including emails, phone numbers, and locations—in the publicly available ShareGPT52K dataset, which contains 50,496 ChatGPT chat histories [60]. Similarly, a report from Cyberhaven revealed that 8.6% of 1.6 million workers pasted confidential data into ChatGPT between its launch and June 2023 [8].

To mitigate the privacy risks associated with users' sensitive personal information in LLMs, several existing research has proposed various *model-centered* approaches to process data before and after the model training to prevent private data leakage [22, 31, 32, 41, 42, 51]. Other research took a *human-centered* approach and assisted users in investigating personally identifiable information leakage [24] or paraphrasing users' sensitive information into less specific terms as they entered it into LLMs [13].

However, these efforts overlook a critical aspect—enhancing users' awareness of how their data is being used by LLMs and helping them understand and reflect on the reasons and risks associated with such usage. This is challenging because, firstly, users tend to prioritize the immediate utility and convenience offered by these systems over the long-term consequences of sharing personal information. The immediate benefits from interacting with LLMs frequently eclipse more abstract concerns such as privacy. This discrepancy leads to an overestimation of the model's safety and an underestimation of its potential to inadvertently expose private data [60]. Additionally, many users lack sufficient privacy literacy, which hampers their ability to comprehend the technical details of data usage [53], model training, and the risks associated with data memorization [19].

This paper fills the gap by focusing on raising users' awareness of the data practices and helping them understand the potential privacy risks in LLM-enabled end-user applications right *at the time of interaction* and *within the context of use.* We aim to help users stay informed of the data practices and potential privacy risks and make informed decisions on whether to proceed with sharing sensitive information (e.g., email addresses, phone numbers, and physical addresses). To do so, we first conducted five co-design workshops with sixteen participants to understand users' privacy needs and explore the design space of privacy awareness tools in LLM-enabled end-user applications. We identified participants' four types of expected privacy information in interacting with LLM-enabled end-user applications, including what data is accessed, who accesses the data, why the data is accessed, and how this access will affect them. They also desired a new approach to help them understand contextual privacy-related information in such applications.

Based on the insights from our co-design workshops, we designed and developed CLEAR, a **C**ontextual **L**LM-**E**mpowered privacy policy **A**nalysis and **R**isk generation tool. CLEAR aims to help users understand the privacy policies of LLMs by analyzing users' current interactions with LLM-enabled end-user applications and the data they are about to share. As shown in Figure 1, the system consists of three parts: 1) detecting the context (i.e., types of sensitive information that the user is about to share); 2) summarizing relevant snippets from the privacy policy of the application ; 3) envisioning potential privacy risks should the sensitive data be shared.

We evaluated its usability and usefulness in two distinct use cases. The first involved a browser plugin for ChatGPT, exemplifying standalone conversational interfaces like Claude and Gemini. In this study, 13 participants engaged in scenarios requiring them to input sensitive data into ChatGPT. The second use case utilized the Gemini addon within Gmail, exemplifying "add-on" LLM tools in existing GUIs, similar to Microsoft Copilot for Office applications. Here, 15 participants used the Gemini plugin in Gmail to reply to or summarize emails containing sensitive information. The results indicated that CLEAR significantly improved the user experience and increased the awareness among participants regarding data practices and privacy risks associated with LLMs. Encouraged by their newfound knowledge of the potential risks and the respective privacy policies, most participants would choose to either delete the sensitive information or substitute it with synthesized fake data, leading to improved privacy management behaviors.

This paper makes the following contributions:

- Through five co-design workshops, we identified user concerns and information needs related to privacy policies and risks when using LLM-enabled end-user applications.
- We introduced CLEAR, a practical tool designed to help users identify privacy risks, understand their implications, and take informed actions *during* their interactions with LLM-enabled end-user applications in a just-in-time manner.
- We explored the dual role of AI in privacy: while it introduces unique privacy risks, it also offers opportunities to improve privacy management.
- We outline design and policy recommendations to enhance privacy awareness, knowledge, and informed decision-making in user interactions with LLMs.

2 Related Work

2.1 Privacy Risks of LLMs

Privacy risks of LLMs are a critical problem that hampers their adoption. Notable incidents, such as the March 2023 ChatGPT bug that leaked user conversation histories and payment information [26], highlight these concerns. Beyond conventional data breaches, three new types of privacy challenges associated with LLMs have emerged: memorization risks, inference risks, and disclosure risks.

Memorization risks. LLMs are trained from vast amounts of data, including user-provided data through interactions. Therefore, these models can memorize and unintentionally reproduce sensitive information from their training data [20, 35]. Even with strict prompt reviews, attackers can exploit models through training data extraction attacks or jailbreaking prompts [28, 47].

Inference risks. Inference risks refer to LLMs' ability to automatically deduce a wide range of personal attributes about individuals from seemingly innocuous text, due to their advanced inference capabilities [30]. For example, phrases like "waiting for a hook turn" can reveal a user is in Melbourne, as "hook turn" is location-specific [47]. While anonymization tools offer some protection, LLMs can still detect subtle cues that reveal personal details.

Disclosure risks. LLMs' human-like interactions foster user trust, increasing the likelihood of unintentional sharing of sensitive information [25, 60, 62]. For example, Zhang et al. [60] found that human-like interactions encourage users to unintentionally share sensitive and personally identifiable information with LLM-based conversational agents. They also noted that users believed they had to "pay the price" of privacy to get benefits from LLMs. This belief, compounded by manipulative interface designs (e.g., dark patterns [34]), amplifies disclosure risks.

These risks stem from users sharing sensitive data with LLMs directly or indirectly. We aim to develop a tool to inform users about privacy policies and potential risks before their data is shared.

2.2 Existing Methods for Protecting User Privacy when using LLM-enabled Applications

Existing methods to protect users' privacy in LLM-enabled applications can be broadly categorized into model-side and user-side solutions. Model-side solutions operate at three stages [46]: pre-processing, training, and post-training. Pre-processing involves data sanitization to remove sensitive information from training datasets, typically using automated techniques like pattern-based parsing [32, 51]. However, defining private information is challenging due to its context-dependent nature and varied formats, making it difficult to guarantee complete sanitization [5]. Training often employs differential privacy, which adds noise to data to prevent individual identification while retaining overall utility [15]. While effective in reducing memorization risks [14, 42], it can degrade model performance and increase computational demands [21]. Post-training methods include filtering outputs with sensitive information [41] and knowledge unlearning, which forces models to forget specific data [22]. However, these approaches face challenges in maintaining output diversity and ensuring consistent performance across use cases [21, 46].

User-side solutions consider user engagement and interaction. Tools like ProPILE [24] let users test for personally identifiable information (PII) leaks in LLMs by probing their own data. However, these tools often lack contextual alignment with actual user scenarios. To address this, self-disclosure abstraction models [13] rephrase sensitive inputs into generalized terms while preserving utility (e.g., "I live in New Mexico" becomes "I live in the Southwest"). Although effective in reducing specificity, these methods do not fully inform users about potential privacy risks or data access.

In summary, existing methods such as privacy-preserving prompting [16] and automated consent mechanisms [43] focus on automating privacy tasks and minimizing user effort. We aim to promote user autonomy by supporting informed, context-aware decisions and enhancing privacy literacy. These approaches =complement each other, that is, existing automated tools can handle straightforward cases, while our approach aims to guide users through complex, ambiguous scenarios that require active intervention.

2.3 Contextual Privacy Policy

Contextual privacy policy (CPP) deconstructs lengthy privacy notices into shorter and context-specific ones that are relevant [18, 56]. Bergmann [3] showed that presenting only relevant information from privacy policies can significantly increase users' privacy awareness. This also aligns with Nissenbaum's theory of contextual integrity [38], which emphasizes that context is critical to determine whether a specific action is a violation of privacy.

CPP has been explored in web and mobile systems. In the general web domain, Ortlof et al. [39] developed a concept showcase of CPP for seven different websites and collected the design implications of CPP. Building on it, Windl et al. [56] presented PrivacyInjector, a more scalable system that can automatically generate and display CPPs. In the mobile domain, Pan et al. [40] proposed SeePrivacy, which integrates vision-based GUI understanding with privacy policy analysis to automatically generate CPPs for mobile applications.

Compared with previous work, our work addresses the unique challenges of implementing CPP for LLM-enabled applications. We not only create CPP for users' sensitive input but also utilize LLMs to highlight potential privacy risks, helping users understand the implications of sharing sensitive information.

3 Co-Design Workshops

To understand user needs and explore the design space of tools to inform participants of privacy risks in LLM-enabled end-user applications, we conducted five co-design workshops[1], including three in-person workshops and two online workshops.

3.1 Participants

Participants are considered "experts of their own experiences [52]" and can contribute their unique perspectives in the design process. We recruited sixteen participants through word-of-mouth and social media, with diverse ages, genders, races, occupations, and previous experiences with LLMs. Table 1 includes a summary of participants' demographic information. The participants were randomly divided into five groups (G1-G5). Each participant was compensated $25 USD for their participation.

3.2 Workshop Procedure

The workshop procedure is inspired by prior work in the privacy literature [34, 57, 58]. Each co-design workshop began with an introductory session, including ice-breaker activities to foster a collaborative atmosphere. Participants then engaged in a discussion on privacy by reading three news articles on incidents relevant to LLMs (e.g., an article entitled "ChatGPT maker OpenAI faces a lawsuit over how it used people's data[2] from the Washington Post)". They identified potential privacy risks and discussed strategies to mitigate them. This initial activity aimed to build a foundational understanding of privacy challenges associated with LLMs.

Participants then engaged in an ideation session to share concerns and opportunities for privacy protection. This transitioned into three activities where they crafted privacy policy elements, assessed risks, and explored privacy incidents through sketching, writing, and discussions (Table 2).

In the first activity, based on an example prompt input into the ChatGPT system, participants designed the content and presentation of retrieved privacy policy snippets that were relevant to sensitive information in example scenarios. Participants could either write descriptions of the interface features or create sketches of the interfaces. The purpose of this ideation activity was to gather user-specific requirements for contextual privacy policies in the context of LLM-enabled applications.

The second activity expanded on the first one by asking participants to examine the prompts, privacy policy snippets, and associated privacy risks. The privacy risks were selected based on Lee et al.'s taxonomy of AI privacy risks [27] and were presented through detailed descriptions and visual aids, including examples of how personal information entered by users could be shared, used for advertising, or stored by service providers. Participants discussed and designed how to organize and present this information to better raise user awareness and support informed actions during user interaction with LLM-enabled applications. The activity specifically focused on the realistic context of use and how such information can fit into the user's overall task that they are trying to achieve through the use of LLM-enabled applications. The goal was to understand whether highlighting LLM-related privacy risks could enhance users' privacy awareness, and if so, how.

The final activity focused on presenting specific privacy incidents, building on the risks identified earlier. Participants designed how real-world examples could highlight privacy implications. The objective was to understand the impact of explicit privacy incident examples on participants' awareness and to determine effective methods for communicating these incidents within the context of LLM application use.

3.3 Data Analysis

All study sessions were audio-recorded and then transcribed for analysis. We followed established open coding procedures [4] to analyze the interview data. Two members of our research team independently coded 20% of the sample, generating a set of preliminary codes using MAXQDA. They compared their codes and reconciled any differences. Using the same codebook, the two researchers coded the rest of the data. In this process, they constantly compared and discussed their codes to ensure full agreement. As the coding process was discussion-based, inter-coder reliability was not necessary [37]. Using the final codebook, we conducted a thematic analysis to identify themes. The complete codebook is provided in the Appendix.

3.4 Key Insights

In this section, we report the key insights we learned from our co-design workshops.

[1] The protocol of the workshop has been reviewed and approved by the IRB at our institution.

[2] https://www.washingtonpost.com/technology/2023/06/28/openai-chatgpt-lawsuit-class-action/

Group	ID	Gender	Age	Ethnicity	Educational Level	Occupation	Used LLM
G1	P1	Female	48	Hispanic or Latino	Bachelor's degree	Office Manager	Yes
G1	P2	Male	39	Black or African American	Bachelor's degree	Maintenance	No
G2	P3	Male	30	Asian and Pacific Islander	Master's degree	Radiation support specialist	No
G2	P4	Female	42	White or Caucasian	Bachelor's degree	Farm owner and writer	Yes
G2	P5	Male	20	Native American or Alaskan Native	Bachelor's degree	Data Analyst	Yes
G2	P6	Female	25	Asian and Pacific Islander	Master's degree	Student	Yes
G3	P7	Female	22	Asian and Pacific Islander	Bachelor's degree	Student	Yes
G3	P8	Female	48	Black or African American	Master's degree	Uber driver	No
G3	P9	Female	29	White or Caucasian	High school graduate	Disabled	No
G3	P10	Male	21	Black or African American	High school graduate	Writer	No
G4	P11	Male	78	White or Caucasian	Doctorate degree	Retired printer	No
G4	P12	Male	35	Black or African American	Bachelor's degree	Programmer	Yes
G4	P13	Female	24	Asian and Pacific Islander	Bachelor's degree	Student	Yes
G5	P14	Male	38	Asian and Pacific Islander	Doctorate degree	Assistant professor	No
G5	P15	Male	41	White or Caucasian	Bachelor's degree	Business management	No
G5	P16	Female	53	White or Caucasian	Master's degree	Librarian	Yes

Table 1: Demographics of participatory design participants

Category	Content
Example Prompt Input into the ChatGPT System	Please help me revise my resume: A diligent and adaptable professional with a passion for problem-solving and a keen eye for detail, seeking to leverage 3-year of experience in finance to contribute effectively to a dynamic team. Eager to apply strong communication and organizational skills to drive positive outcomes in a collaborative work environment. Contact: 987-123-4567.
Related Privacy Policy Snippets	*"Category of Personal Information: Identifiers, such as your name, contact details, IP address, and other device identifiers."* *"Disclosure of Personal Information: We may disclose this information to our affiliates, vendors, and service providers to process in accordance with our instructions."* *"As noted above, we may use Content you provide us to improve our Services, for example, to train the models that power ChatGPT."*
Potential Privacy Risks	**Model Memorization (Secondary Use):** LLMs can memorize and potentially regurgitate snippets of training data. It can lead to data leakage, where personal information becomes accessible through the model's responses. **Targeted Advertising (Increased Accessibility):** Once your contact information is shared, it could be used for targeted advertising, which might be intrusive. **Loss of Anonymity (Identification):** Sharing your contact information can link your identity to specific activities, opinions, or interactions with the model, which might otherwise remain anonymous.

Table 2: Example materials used in the co-design workshop. Note that the parentheses in the Potential Privacy Risks row highlight the types of privacy risks based on a taxonomy of AI privacy risks [27].

3.4.1 KI1: Concerns About Privacy Policies in AI Contexts.

Participants expressed concerns about four primary areas: the types of data accessed, the entities that access the data, the reasons for data access, and the potential impacts of such access. The context of AI makes it challenging for them to fully understand these aspects.

Participants were particularly uncertain about the specific types of personal information collected, including concerns about access to payment transactions, contact details, frequently used services, and online behavior tracking. For example, one participant in G5 in our study said "*I do look for things like what are they going to be looking at in my device.*" The sheer volume and variety of data that LLMs can access increases the risk of unintentional data breaches and misuse. Users may not be fully informed about the extent of the data being accessed, leading to a lack of informed consent. Additionally, participants were concerned about who would use the sensitive data, emphasizing the need to understand if their data is shared with any third parties. The complexity of AI systems often necessitates data sharing between different entities, increasing the risk of unauthorized access and data misuse. For example, a participant in G3 hypothesized that "*[The sensitive information] can be given to other companies and can be shown on other search results.*"

Participants also questioned the purpose of the data collection. Some of them wondered if their data was used for improving services or for targeted advertising, which blurred the lines of user consent, as it was not clear to users what they had consented to. Furthermore, another significant concern was the potential impacts on users' privacy. A participant in G1 said "*I really want to know if my personal health information will be shared with a credit card company or a life insurance company... I think that it would be very helpful for users if they had such information.*"

3.4.2 KI2: Expectations for Contextual, Simplified, and Focused Privacy Information. The participants expressed a desire for contextual, concise, and easy-to-understand privacy information, highlighting the importance of clearly communicating the reasons and consequences of data use. For instance, participants in G4 suggest adding tooltips to show definitions of words in the policy ("*if I am not really sure what this means, I can click on it to quickly get a definition for what that means.*"), "*visualizing important information on one page while providing links to additional details*", and using simple language to summarize the information concisely. Participants emphasized that they "*don't want to be exposed to all the information*", instead they expect "*one page to include only the most important (privacy) information.*" Furthermore, participants highlighted the need for privacy policies and risk alerts to be relevant to the context in which they are engaged. They proposed features like highlighting sensitive information, summarizing pertinent policy sections, and outlining potential privacy risks in a clear manner For example, one participant in G5 expected to have a tool to "*highlight the key permissions you are granting in bullet points which... you should be aware of at a glance, ensuring you know exactly what you're consenting to.*" Such brief and in-context privacy notices could also enhance participant understanding and trust towards LLMs.

Regarding privacy incidents, while participants acknowledged the value of presenting information about privacy breaches, they often do not engage with such content as they lack the time to read the lengthy articles thoroughly. Additionally, if the news content does not directly relate to the types of sensitive information they are concerned with, their motivation to read decreases significantly. This disengagement greatly diminishes the potential impact of such example incidents on enhancing their privacy awareness.

3.4.3 KI3: Desire for Enhanced Data Protection and Increased Privacy Awareness. Participants reported that they are already actively engaged in various methods to protect their privacy, such as data sanitization (e.g., "*providing some fake or random location*" and "*use different fake names on different websites*"). They also adopt secure browsing practices, including using unique passwords for each site, employing stricter authentication measures, and opting for more secure browsers. They further limit data sharing by controlling the information they disclose, storing private data locally, and avoiding sharing sensitive information with untrusted applications or individuals.

Beyond these protective measures, participants expressed a desire to improve their privacy awareness, such as regularly reviewing app permissions and learning about the privacy risks. However, they seek a more profound understanding of privacy policies. For example, a participant in G2 anticipated that the existing online service to "*show more transparent connection between the features and the data, like how my data supports what feature.*" Participants also emphasized the importance of recognizing the value of their data and the availability of resources to improve their privacy literacy. For instance, one participant in G5 emphasized "*showing the economic value is so important because my data are so valuable.*" While another noted, "*I want an interactive tutorial (to) help you go through all the features and like potential privacy risk you may have in the app.*"

3.4.4 KI4: Aspirations for AI to Enhance Privacy Control and Protection. Despite being aware of the privacy risks associated with AI, participants still see it as a valuable tool for enhancing their privacy protection. They proposed several ways in which AI could be leveraged to safeguard their personal information: Specifically, participants suggested several ways AI can enhance privacy protection: inspecting sensitive information in participant inputs to prevent unintentional sharing, having language models unlearn data upon request, using AI to monitor participants and defend against hacking, and detecting risky data traffic to block it. For example, a participant in G5 expected that "*there should be another AI supervisor that can monitor these events and report to police.*" Instead of relying on stricter privacy policies or new privacy regulations and laws, participants want AI to help them directly, indicating a desire for more engagement and control to protect their privacy through AI technology. Additionally, they emphasized the need for AI to assist participants in recognizing and managing the data they share and allowing participants to accept, deny, or modify information before submission. For example, a participant in G1 mentioned "*at the end of the chat, the AI can give the user a selection of whether they want their data to be shared.*"

3.5 Design Goals

The following key design goals were developed based on insights from our formative study and are instrumental in guiding the design of our solution:

- **DG1: Enhance Understanding of Data Access and Usage.** It should help participants understand (1) what sensitive data is accessed by LLMs, (2) who accesses it, (3) why it is accessed, and (4) how it affects them. It should provide clear explanations of LLMs' data sharing practices and their implications for the users (**KI1** in Section 3.4.1).
- **DG2: Develop Contextual and User-Friendly Privacy Policies.** It should develop privacy policies that are not only succinct but also contextual and easy to comprehend. It should summarize important context-related information in the privacy policy into a single page that serves as a guide to help users access more detailed information if needed (**KI2** in Section 3.4.2).
- **DG3: Leverage AI to Manage Data Sharing and Enhance Privacy Literacy.** It can use AI to inspect sensitive information and prevent unintentional sharing. Furthermore, it can leverage AI to improve participant understanding of privacy policies by extracting contextual privacy policies and generating potential privacy risks (**KI3** in Section 3.4.3).
- **DG4: Promote Participant Control and Engagement in Privacy Management.** It should highlight key permissions and potential privacy risks to ensure that participants

are fully aware of what they are consenting to. This goal seeks to enhance participants' engagement and control in safeguarding their privacy, fostering a more informed user base (**KI4** in Section 3.4.4).

4 System Design

4.1 Overview

To address the outlined design goals, we introduce CLEAR: a **C**ontextual **L**LM-**E**mpowered privacy policy **A**nalysis and **R**isk generation tool. As illustrated in Fig. 2, CLEAR consists of three main components: detecting the context and type of sensitive information from user inputs, retrieving and summarizing relevant privacy policy snippets of the underlying LLM-enabled application, and identifying potential privacy risks associated with sharing such information with the underlying LLM-enabled application.

4.2 Example Use Case

We illustrate the functionality of CLEAR through two specific example use cases: direct interactions with LLMs through conversational interfaces (e.g., ChatGPT) and interactions with LLM-based add-ons in existing graphical user interface (GUI) applications (e.g., Gemini plugin in Gmail).

In the first scenario (Fig. 1), a user, Alice wants to ask ChatGPT to help her write an email to follow up with a client regarding an unpaid invoice. After detecting sensitive information in Alice's input, CLEAR alerts her to the specific types of sensitive data involved, along with relevant privacy policy snippets and potential risks.

In the second scenario (Fig. 3), another user, Randy would like to ask the Gemini plugin in Gmail to summarize an email. After identifying sensitive information in the email, CLEAR notifies Randy about the specific categories of sensitive data detected, providing related privacy policy excerpts and highlighting potential risks. The tool enables users to identify privacy risks, understand their implications, and make informed decisions in real time while interacting with LLM-enabled applications.

- *Indicating sensitive Information*: Alice asks ChatGPT to help her draft an email and enters the following prompt: "Help me write an email. My email is abc@gmail.com and my contact is 1234567890. I need to politely follow up on an unpaid invoice, reminding the client of the due date and offering assistance if needed." As Alice types, CLEAR detects sensitive information (i.e., personal email and phone number) and displays a notification with the corresponding types of sensitive information: "You input sensitive information: email, contact."(Fig. 1a)
  Similarly, as Randy enters "summarize this email" in the Gemini plugin, CLEAR detects sensitive information (i.e., phone number and physical address) and pops up a notification: "Sensitive information in this email: address, phone number."(Fig. 3a)
- *Display relevant privacy policy snippets and potential privacy risks*: Alice clicks the "Know more" button in the CLEAR pop-up to learn about privacy policy snippets related to the detected sensitive information and the potential privacy risks of sharing sensitive information with LLMs (Fig. 1b). Each type of sensitive information corresponds to specific policy snippets and associated risks. For example, contact information is mentioned at least three times in ChatGPT's privacy policy. In the policy snippets section (Fig. 1b(1)), Alice learns about the data accessors of this type of sensitive information and their reasons for accessing it. She clicks "Refer to original text" to view detailed excerpts from the full privacy policy. In the potential risks section (Fig. 1b(2)), Alice gains insights into the risks of sharing each type of sensitive information with LLMs. By clicking "Detail explanation," she can access an in-depth description of the privacy risks associated with each category. This information helps Alice make a more informed decision about whether she should share sensitive information with ChatGPT.
  Similarly, Randy accesses policy snippets and potential risks of the Gemini plugin by clicking "Know more," seeing who can access his data and evaluating any potential risks, which aids his decision-making.

4.3 Key Features of CLEAR

4.3.1 *Identifying sensitive information.*

Sensitive information typically includes personally identifiable information (PII) such as email addresses, phone numbers, and physical addresses. These data types have previously been exposed in conversation histories between users and ChatGPT [60]. The first component of CLEAR focuses on identifying sensitive information from user input, helping users become aware of privacy risks, and motivating them to exercise greater caution when interacting with LLMs.

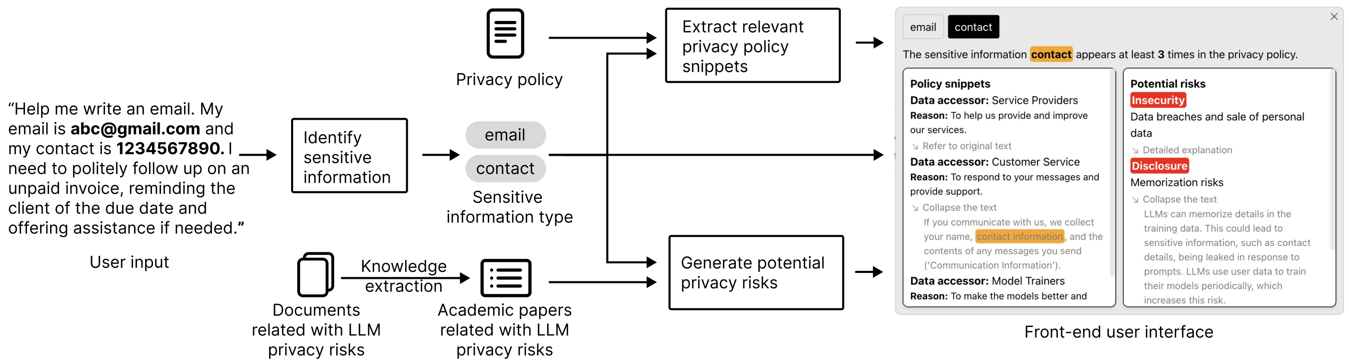
To ensure that users do not directly share their sensitive information with LLMs while using CLEAR, we first use Microsoft Presidio[3] to identify the types of sensitive information in the user's input. In the case of direct conversational interfaces for LLM (e.g., ChatGPT), the user's input includes the prompt sent to the LLM. For LLM add-ons (e.g., the Gemini addon for Gmail), the user input also includes the user's private data in the context (e.g., the content of the underlying email). CLEAR passes only the types of information, rather than the actual information itself, to the LLM for the subsequent extraction of contextual privacy policies and the generation of potential privacy risks, as detailed below.

4.3.2 *Extracting relevant privacy policy snippets.*

According to **KI2**, which highlights the need for users to access only the most relevant information for their ongoing tasks, this component of CLEAR focuses on extracting relevant snippets from the full privacy policy of the corresponding LLM based on the types of sensitive information identified in the previous step. This is to ensure that the information provided aligns with the users' immediate context and needs.

To achieve this, we use few-shot learning techniques [5], which enable our system to learn effectively from a limited set of examples and generalize this learning to identify relevant sections of privacy policies. This approach allows CLEAR to quickly adapt to different privacy policies and pinpoint the most crucial snippets. Specifically, these snippets detail who has access to the sensitive information and for what purposes, tailored to the types of information users have entered (**KI1**). This method speeds up the process and enhances the contextual understanding for users, helping them better

[3] Microsoft Presidio: https://microsoft.github.io/presidio/


"Help me write an email. My email is **abc@gmail.com** and my contact is **1234567890.** I need to politely follow up on an unpaid invoice, reminding the client of the due date and offering assistance if needed."
User input
Identify sensitive information
email
contact
Sensitive information type
Privacy policy
Extract relevant privacy policy snippets
Documents related with LLM privacy risks
Knowledge extraction
Academic papers related with LLM privacy risks
Generate potential privacy risks
Front-end user interface

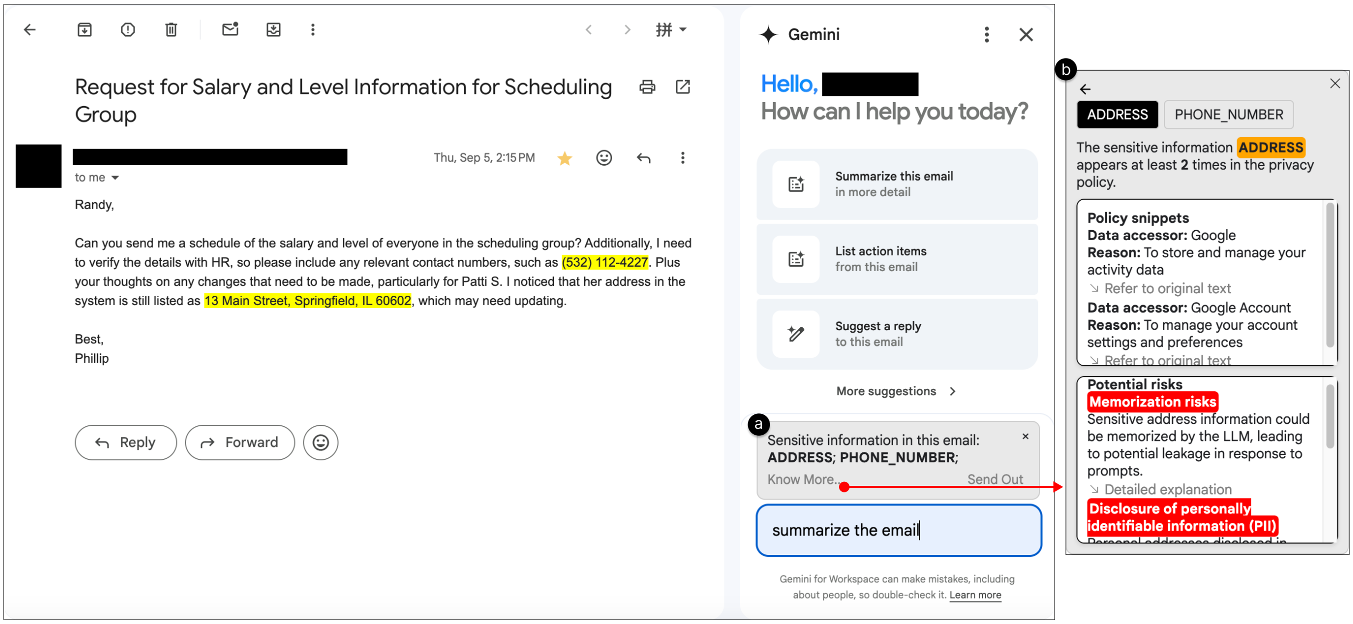
email contact

The sensitive information **contact** appears at least **3** times in the privacy policy.

Policy snippets
Data accessor: Service Providers
Reason: To help us provide and improve our services.
↘ Refer to original text
Data accessor: Customer Service
Reason: To respond to your messages and provide support.
↘ Collapse the text
If you communicate with us, we collect your name, contact information, and the contents of any messages you send ('Communication Information').
Data accessor: Model Trainers
Reason: To make the models better and

Potential risks
Insecurity
Data breaches and sale of personal data
↘ Detailed explanation
Disclosure
Memorization risks
↘ Collapse the text
LLMs can memorize details in the training data. This could lead to sensitive information, such as contact details, being leaked in response to prompts. LLMs use user data to train their models periodically, which increases this risk.

Figure 2: System flow of CLEAR



Request for Salary and Level Information for Scheduling Group

[redacted] [redacted] Thu, Sep 5, 2:15 PM
to me

Randy,

Can you send me a schedule of the salary and level of everyone in the scheduling group? Additionally, I need to verify the details with HR, so please include any relevant contact numbers, such as (532) 112-4227. Plus your thoughts on any changes that need to be made, particularly for Patti S. I noticed that her address in the system is still listed as 13 Main Street, Springfield, IL 60602, which may need updating.

Best,
Phillip

Reply Forward

Hello, [redacted]
How can I help you today?

Summarize this email in more detail

List action items from this email

Suggest a reply to this email

More suggestions >

Sensitive information in this email: **ADDRESS; PHONE_NUMBER;**
Know More... Send Out

summarize the email

Gemini for Workspace can make mistakes, including about people, so double-check it. Learn more

ADDRESS PHONE_NUMBER

The sensitive information **ADDRESS** appears at least **2** times in the privacy policy.

Policy snippets
Data accessor: Google
Reason: To store and manage your activity data
↘ Refer to original text
Data accessor: Google Account
Reason: To manage your account settings and preferences
↘ Refer to original text

Potential risks
Memorization risks
Sensitive address information could be memorized by the LLM, leading to potential leakage in response to prompts.
↘ Detailed explanation
Disclosure of personally identifiable information (PII)
Personal addresses disclosed in

Figure 3: The user interface of CLEAR in the Gmail Context. In this example, a user attempts to use the Gemini plugin in Gmail to summarize an email containing sensitive information, such as an address and phone number. CLEAR detects and highlights the sensitive information in a pop-up(a). When the user clicks "Know more...", the pop-up expands to show relevant privacy policy excerpts from Gemini, along with potential privacy risks identified by the LLM(b). The black rectangles in the image are added to the screenshot to obfuscate the user's information from being disclosed in the figure.

comprehend the potential privacy risks associated with entering their data into LLMs.

> **Prompt for generating privacy policy snippets**
>
> Given the privacy policy: {policy}, and the input sensitive information: {sensitive information}, cite the sentences in the privacy policy that are semantically related to the input sensitive information only when using the LLM service rather than creating an account. List who will access the data (service provider and third parties) and the reason for access. Each data accessor should be different. The language and vocabulary should be easy to understand for kids in elementary school.

4.3.3 Generating potential privacy risks. The final component of CLEAR involves generating potential privacy risks associated with sharing sensitive information with LLMs. We use a chain-of-thoughts approach [54], which involves sequentially reasoning through the potential implications of data sharing. This method is structured as follows:

- **Knowledge Extraction**: CLEAR first uses LLMs to extract structured knowledge from academic papers that discuss LLM privacy risks. Table 3 shows a list of academic papers that we used for knowledge extraction. These papers were initially identified through an automated search process using Google Scholar using the keywords `privacy risk AND`

Paper Name	Citation
A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly [59]	90
Risk taxonomy, mitigation, and assessment benchmarks of large language model systems [9]	16
Security and privacy challenges of large language models: A survey [12]	11
Deepfakes, Phrenology, Surveillance, and More! A Taxonomy of AI Privacy Risks [27]	4

Table 3: An Example paper list for generating privacy knowledge

`(large language model OR LLM)`. After retrieving the relevant papers, we prioritized those that included surveys or user studies related to LLM and privacy risks. The knowledge extraction process was semi-automated: LLMs were used to analyze the text of these papers, extracting key insights and structuring them into predefined categories such as privacy risks and their causes, influences, and evidence. Then, these structured outputs were manually verified by experts for accuracy and relevance. This combination of automated extraction and human validation contributes to the generalizability of the system by ensuring that the generated risks are both comprehensive and grounded in robust, well-researched sources. The structured knowledge generated from this process is detailed in the Appendix, where examples of the extracted risks are provided.

- **Prompt Generation**: The next step involves using the extracted knowledge to craft prompts that guide the LLM in generating specific privacy risks. These prompts incorporate context-specific details based on the sensitive information previously identified, ensuring that the risks generated are relevant to the user's situation.
- **Risk Generation**: Using the prompts, the LLM envisions and articulates potential privacy risks based on the context of use and the types of private data included in user inputs. This step involves outlining scenarios where sensitive information could be misused or inadvertently exposed, providing users with a comprehensive view of the risks involved. Building on the taxonomy of AI privacy risks proposed by Lee et al. [27], we tag each generated privacy risk with its corresponding risk category. Inspired by **DG2**, each privacy risk has a brief summary and an expandable detailed explanation. This method ensures users receive a clear, context-aware understanding of the privacy risks they face when interacting with LLMs.

Prompt for generating potential privacy risks

You are acting as a privacy expert, and you know the knowledge about privacy risks in the context of LLM: {structured knowledge}, Given the privacy policy: {policy}, Return the most critical 3 LLM privacy risk with detailed consequence as evidence (e.g., memorization risks). The risks should be related to the sensitive information: sensitive information. Also, return risk type and related privacy knowledge.

4.4 Implementation

The front-end chrome extension of CLEAR is implemented in Node.js[4] and packed by Webpack.js[5]. Its back-end server was developed using the FastAPI framework[6]. CLEAR uses the Microsoft Presidio library[7] to identify instances of private information in user inputs and infer their types. CLEAR uses the LangChain[8] and Pydantic[9] libraries to extract contextual privacy policies, summarize structured knowledge from past papers related to AI privacy risks, and generate potential privacy risks.

5 User Studies

We conducted two user studies on two use cases separately[10] to evaluate the usability and effectiveness of CLEAR, with a specific focus on how CLEAR influenced users' understanding and actions regarding privacy in LLM-enabled end-user application. The first case aimed to assess how CLEAR aids users in managing their privacy while using LLMs in a dialogue-based context in ChatGPT. The second case aimed to explore CLEAR's effectiveness in environments where LLMs augment traditional interfaces while using Gmail augmented by a Gemini plugin. These two use cases allowed us to gain comprehensive insights into how CLEAR functions across different environments, characterized by varying levels of privacy risks and user interaction patterns.

5.1 Participants

For the first study with ChatGPT, we recruited 13 participants, and for the second study using the Gemini plugin in Gmail, we recruited 15 participants. All were recruited via word-of-mouth and social media (e.g., Facebook), and each completed a pre-screening survey to provide demographic details such as age, gender, location, occupation, education, and ethnicity, to ensure a diverse group of participants. The first study had four females and nine males aged 25 to 48 (see Table 4), and the second study had seven females and eight males aged ranging from 23 to 58 (see Table 5). Both studies were conducted virtually using Zoom, with each session lasting approximately one hour. For their participation, each individual received a $25 gift card.

[4] https://nodejs.org/en
[5] https://webpack.js.org/
[6] https://fastapi.tiangolo.com/
[7] https://microsoft.github.io/presidio/
[8] https://www.langchain.com/
[9] https://docs.pydantic.dev/latest/
[10] The study protocol was reviewed and approved by the IRB at our institution.

ID	Gender	Age	Ethnicity	Educational Level	Occupation
P1	Female	42	White or Caucasian	Bachelor's degree	Writer
P2	Male	25	Black or African American	High school graduate	Student
P3	Male	28	Hispanic or Latino	High school graduate	Project Manager
P4	Male	38	Asian and Pacific Islander	Doctorate degree	Teacher
P5	Female	48	Hispanic or Latino	Bachelor's degree	Office Manager
P6	Female	27	Black or African American	Bachelor's degree	GIS Analyst
P7	Male	25	Black or African American	Bachelor's degree	Student
P8	Male	36	White or Caucasian	Master's degree	Teacher
P9	Male	27	Black or African American	Bachelor's degree	Architect
P10	Female	25	Asian and Pacific Islander	Master's degree	Student
P11	Male	36	White or Caucasian	Doctorate degree	Teacher
P12	Male	26	Black or African American	Bachelor's degree	Student
P13	Male	35	Black or African American	Bachelor's degree	Computer programmer

Table 4: Participant Demographics in Case Study 1

ID	Gender	Age	Ethnicity	Educational Level	Occupation
P1	Female	29	Black or African American	Bachelor's degree	Teacher
P2	Male	28	White or Caucasian	Master's degree	Web developer
P3	Female	23	Black or African American	Bachelor's degree	Student
P4	Male	24	Black or African American	Bachelor's degree	Student
P5	Female	58	White or Caucasian	High school graduate	Creative writer
P6	Female	33	Black or African American	Bachelor's degree	Landscaper
P7	Male	23	Black or African American	High school graduate	Freelancer
P8	Male	30	White or Caucasian	Master's degree	Landscape architect
P9	Male	26	Black or African American	Bachelor's degree	Software Engineer
P10	Male	29	Hispanic or Latino	Bachelor's degree	Data Analyst
P11	Male	28	Black or African American	Bachelor's degree	Developer
P12	Female	26	Asian and Pacific Islander	Master's degree	Finance intern
P13	Male	40	Asian and Pacific Islander	Doctorate degree	Professor
P14	Female	27	Black or African American	Bachelor's degree	Data Scientist
P15	Female	30	Black or African American	Bachelor's degree	UX Designer

Table 5: Participant Demographics in Case Study 2

5.2 Study Procedure

5.2.1 Use case of ChatGPT. We employed a within-subjects study design. Each participant was asked to enter two prompts to ChatGPT, one before using CLEAR and one after using it, and then compared their experiences. Below are the detailed procedures.

Each session contained the following parts. After obtaining consent, we introduced the team and study objectives, followed by a pre-intervention assessment. We asked the participants to enter one of the following two prompts in ChatGPT, and then share their perceptions about who would access what information, why, and how it would affect them. We selected these two prompts because both represented a typical scenario of interacting with LLMs that involve user inputs containing personally identifiable information. The prompts were counterbalanced to mitigate biases.

After the pre-intervention assessment, we introduced CLEAR to our participants and asked them to use it freely for five minutes to explore its features and get familiar. Then, participants entered the other prompt to ChatGPT with CLEAR enabled. At this time, the CLEAR interface was triggered as the participants entered the prompts and provided additional information as described in Section 4. Participants reviewed this information and shared their updated perceptions.

Following both conditions, we conducted interviews to compare their experiences with and without CLEAR, focusing on differences and the impact of policy snippets and privacy risk details on their understanding and actions.

Finally, participants completed a post-study questionnaire, including the System Usability Scale (SUS) and questions on overall experience, content quality, and consistency.

The two prompts used in case study 1

Prompt 1: I need help revising my resume profile. I'm Michael Brown, reachable at michaelbrown@gmail.com and (555) 123-4567, residing at 789 Pine Road, River City, CA 90210. My current profile states: "Dedicated marketing professional with a proven track record in digital marketing strategies, SEO, and content creation. Adept at driving brand growth and engagement.

Prompt 2: Can you help me enhance my resume profile? My details are: Sarah Johnson, sarahjohnson@gmail.com, (444) 567-8901, living at 321 Birch Lane, Capital City, TX 73301. My profile currently reads: "Results-oriented project manager with extensive experience in leading cross-functional teams, managing budgets, and ensuring timely project delivery.

5.2.2 Use case of the Gemini plugin in Gmail. We implemented a similar within-subjects design for this study that focused on participants' experiences with the Gemini plugin in Gmail. Each participant reviewed six emails—three without using CLEAR and three with it. The detailed process is outlined below.

Dataset. We randomly selected six emails from the Enron public email dataset[11] as the dataset for the case study. Since LLM-generated synthetic data is widely used for data augmentation [45], and to ensure that the selected emails contained sensitive information, we used GPT-4 to synthesize the sensitive information and integrated it into the selected six emails.

Each session contained the following parts. After the consent, we began with introductions, where the research team and the objectives of the study were presented. The second part was a pre-intervention assessment. We asked the participants to review and summarize the content of three emails randomly selected from the six emails in the dataset. Based on this experience, we asked participants' perceptions about who would access what information, why, and how it would affect them. To control for potential biases, we counterbalanced the order in which the participants encountered the emails.

After reviewing the first three emails, participants were introduced to the CLEAR tool and given five minutes to familiarize themselves with it. They then reviewed the remaining three emails with CLEAR enabled and summarized their content by using the Gemini plugin. CLEAR's interface was activated as the Gemini plugin processed commands, allowing participants to see the privacy-related details provided. Following this, participants were again asked to reflect on their perceptions of who could access the sensitive information, why, and the associated risks.

Similar to the first case study, after the two conditions, we interviewed the participants, asking them to compare their experiences with and without CLEAR and explain the differences, if any, and how the policy snippets and potential privacy risks impacted their understanding and potential actions. Finally, participants completed a post-study questionnaire that included the System Usability Scale (SUS) and questions about their overall user experience and the quality and consistency of the contents presented to them.

A sample email used in case study 2

Hello.
I gave the Matagorda County Fair & Livestock Association 2050 USD using my credit card number 1234-5678-9012-3456, and filled out the form for the match. My billing address is 1234 Elm Street, Springfield, IL 62704. Please advise if this has been paid, as they are sending me a bill for the 2050 USD, plus a finance charge. You can reach me at 555-123-4567 if you need more information.
Thank you,
Kay Mann

5.3 Data analysis

We followed the same approach for the data analysis for the results of both case studies. For the user rating data (i.e., SUS scores), we computed the mean and standard deviation for each statement. For the interview data, two coders independently reviewed and coded the transcripts, following the method described in Section 3.3. After completing their individual coding, the coders iteratively refined the codebook through discussion, identifying recurring themes. These themes included changes in privacy perceptions after using CLEAR, reactions to the prompts, and suggestions or feedback on the tool. The detailed codebooks used and refined during this analysis are available in the Appendix. This structured data analysis approach ensured that we could systematically capture and interpret both the quantitative and qualitative feedback on CLEAR, providing a comprehensive understanding of its impact on users' privacy awareness and behavior.

6 Results of Case Study 1

6.1 System Usability

The evaluation of CLEAR in the context of ChatGPT, based on the results of the System Usability Scale(SUS) questionnaire [2], indicates a generally positive user experience, as shown in Fig 4. Participants rated each item on a 5-point Likert scale. They find the system easy to use and well-integrated, with most expecting that it would be quickly learned. The confidence in using the system is high, and the information provided is deemed useful. Additionally, policy snippets and privacy risks are considered appropriate and well-matched with the sensitive information. Participants did not perceive the system as complex, cumbersome, or annoying, nor did they feel a need for technical support.

6.2 Impact on Participants' Understandings and Actions Regarding Privacy in LLMs

Our study results suggest that using CLEAR (1) increases participants' awareness of data practices in LLMs, (2) raises participants' awareness of new privacy risks, and (3) leads to participants' intention for behavior changes. We present the details below.

6.2.1 Increase participants' awareness of data practices in LLMs (P1, P2, P4, P6, P7, P9, P11, P12). By comparing participants' reactions

[11] https://www.cs.cmu.edu/~enron/

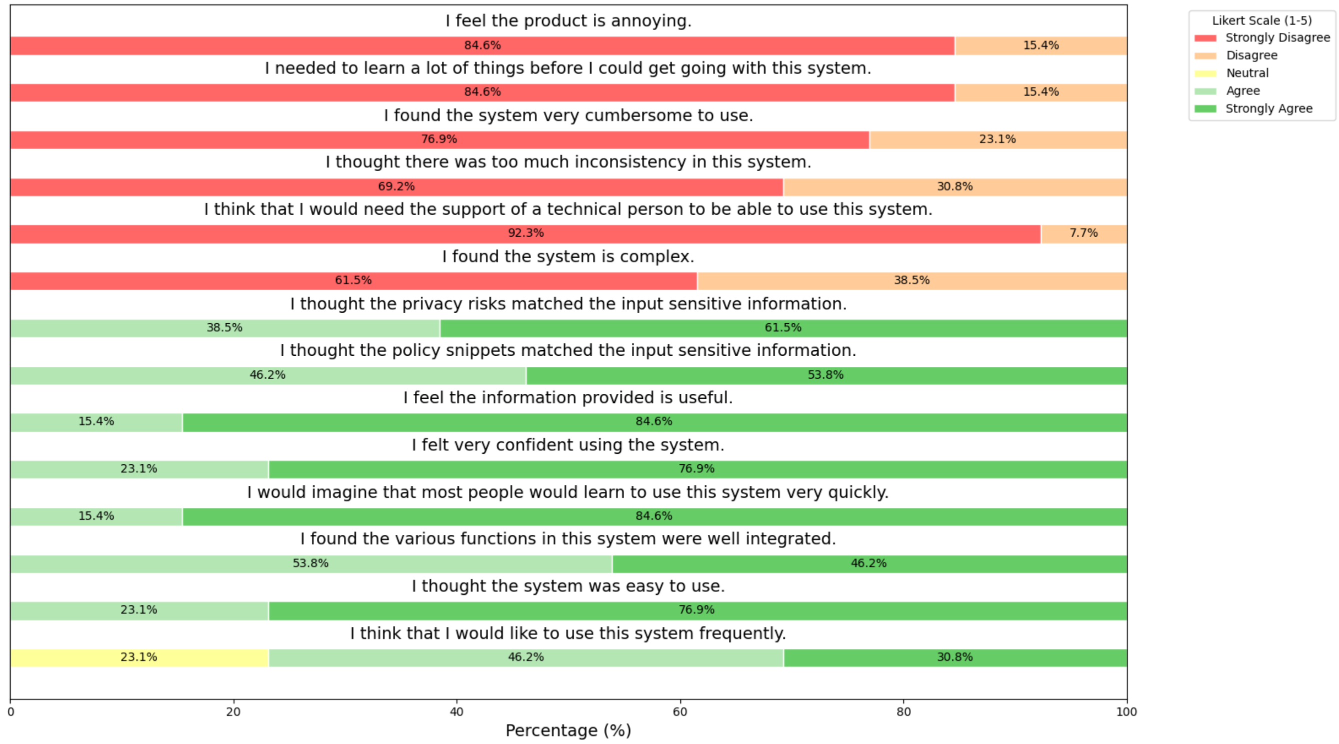


Figure 4: Result of Post-study Questionnaire for the Case Study 1.

to the first prompt before they used CLEAR and the second prompt after they used it, we observe that CLEAR enhances participants' understanding of various aspects of their data. Table 6 provides an overview of such comparisons. More specifically, before using CLEAR, participants tended to believe that legitimate access to their private data was limited to the companies developing large language models, as shown in Table 6. However, after using CLEAR, participants quickly realized that there are more data accessors, including vendors, affiliates, and government agencies. For instance, P2 expressed surprise by saying, "*Honestly, I did not know that even law enforcement has access to all this information. I was surprised.*" This statement underscores a significant gap in their initial understanding, revealing a lack of awareness regarding who can access their information.

Meanwhile, participants also showed a deeper understanding of the reasons why their sensitive information is accessed. For example, P9 initially thought that data breaches were primarily due to hackers: "*I think the hackers will have access to this information because this is personal information that should not be put on a public website unless you are assured that your details are safe.*" Later, they understood that many more entities could access their data, which amplified the risks for data breaches: "*Vendors and service providers, OpenAI, customer service, training team, analytics team, affiliates are also gonna see your information. I think this is because these entities work on the ChatGPT platform. So they're gonna see the information. The best way to curb this is to not share your personal information at all.*" This newfound clarity extended to data access purposes, and participants became more aware of their information being used not only for service improvement but also for training models and other purposes.

6.2.2 Raise participants' awareness of privacy risks in LLMs (P1, P2, P3, P7, P9). The use of CLEAR assists participants in understanding and clarifying possible privacy risks when using LLMs. Initially, participants only had a general understanding of privacy risks, but the tool provided detailed insights that clarified these risks. P1, for example, noted that CLEAR highlighted specific areas where their data was at risk, allowing a better understanding of potential privacy issues. She said "*The tool showed me exactly the areas where it was at risk and put it into focus so that I could have a better understanding of it. [Initially,] I had a general vague fuzzy idea of what the risk was and then the tools showed me exactly [the risks]...The tool removes the vagueness in my head of what LLM could actually do.*" Similarly, as shown in Table 6, P4 initially believed that data access was only for service improvement, but later recognized the potential for commercial use of the data and sharing of the data with third-parties, stating, "*My information could be sold. People could find a lot of my information. My information might end up being saved and we're places where it shouldn't be.*"

6.2.3 Lead to an intention for behavior change (P4, P6, P9). The increased awareness and understanding fostered by the use of CLEAR led our participants to consider changing their behavior to better protect their privacy. P6, for example, concluded that using LLMs necessitates more cautious sharing of personal information: "*The*

Questions	**Initial Assessment**	**Post-tool Assessment**
WHAT are sensitive information	P1:“*ChatGPT will have access to address, email, phone, professional status.*”	P1:“*ChatGPT stores the sensitive information given by the user (email, address, contact info).*”
WHO will access the sensitive information	P2: “*The researchers at OpenAI will access this information.*” P5: “*I doubt ChatGPT may share my contact information with different marketing agencies.*” P6: “*The AI will access the profile.*”	P2: “*Service providers, government agencies, affiliates will access this information.*”. P5: “*OpenAI will share the data with vendors, service providers, and law enforcement agencies.*” P6: “*There are service providers, and multiple affiliates, that will access this information.*”
WHY does the entity want to access the sensitive information	P1:“*For the owner of the AI model, there is no reason to access, unless they would be asked to by an authority with greater credentials.*” P9:“*I think the hackers will have access to this information because this is personal information that should not be put on a public website unless you are assured that your details are safe.*”	P1:“ *Law enforcement agencies could have an overview of what the user has been doing without him/her being aware of the information disclosed.*” P9:“*The government authorities will be able to access the data, so I think it is wise to not put your personal information on there. Vendors and Service providers, Open AI, Customer service, Training team, Analytics team, Affiliates are also gonna see your information. I think this is because these entities work on the ChatGPT platform. So they’re gonna see information. The best way to curb this is to not share your personal information at all.*”
HOW will the access of sensitive information affect the user	P4:“*The access of the data will help the user have a more relevant and accurate answer because the modal will know more about you and be able to better help and assist you.*” P6:“*The AI is accessing this information in order to help the user revise their resume profile.* ” P7:“*The user will not have control of how the AI will use the data after it has provided the resume.*”	P4:“*The access could lead to information being sold or mishandled but it will also allow for a more accurate response.*” P6:“*There are no promises that all of this will not be shared or breached. The user should reconsider providing the name, address, and email address with the AI and just use the one sentence summary that he/she is requesting help with.* ” P7:“*The user will be affected in that his/her information can be exposed(memorized) and also disclosed to other third party users. The user may also experience data breach.*”

Table 6: Comparison of Initial and Post-tool Assessments

user should reconsider providing the name, address, and email address with the AI and just use the one-sentence summary that he/she is requesting help with...I need to take steps to make sure that I’m at least not providing over information about myself other than basic things like updating a resume”. P9’s statement further highlighted the importance of not sharing personal information publicly, “*The government authorities will be able to access the data, so I think it is wise to not put your personal information on there.*” This collective change in behavior reflects the participants’ intentions to adopt more privacy-conscious practices as a result of their improved understanding of LLM privacy risks.

6.3 Participants’ Reactions to the Disclosure of Sensitive Information

6.3.1 Most participants want to remove sensitive information from prompts or replace it with fake data (P1, P2, P3, P5, P6, P7, P8, P11, P12). Our participants expressed a strong desire to remove sensitive information from their prompts or substitute it with fake data. P1 explicitly stated their preference for removing all personal information: “*I think I would remove all the all the personal information.*” Similarly, P6 emphasized the importance of excluding personal details such as name, email address, telephone number, and home address: “*I just removed like my name my email address, my telephone number. And my home address from the prompt.*” Additionally, P5 highlighted an alternative approach by suggesting the use of fake contact details: “*I may like disguise him by providing fake email and fake phone number. I will not provide these details and only give a generic name, email, and phone number.*” These responses indicate a common practice among our participants to ensure their privacy and security by omitting or obfuscating sensitive data in their prompts. Some participants even suggested that CLEAR could directly provide solutions for them, as P4 mentioned, “*I would like to know more like not only the information they present but also some solutions that I can choose.*”

However, several participants chose to proceed with sending the sensitive information without any modification. For instance, P13 stated, *It wouldn’t stop me from using ChatGPT. I’ll send it out.*”

They believed that this information could still be obtained by other applications in other ways, regardless of whether it was shared here or not. This opinion was further explained by P5: *Because I already know that they [other online services] have access to this information. These days, your email and your phone number or your mailing address are some of the easiest things to get. So, I'm okay that if they are using so."*

6.3.2 Participants expect CLEAR to identify more types of sensitive information and define the scope of "sensitive" clearly (P1, P2, P4, P5, P5, P7, P12). Participants showed a strong desire for CLEAR to identify more modalities and types of sensitive information. P2 expresses a need for the system to identify not only text but also sensitive information in images and PDF files: "*I want it can not only identify this text but also the things in the image and PDF files.*" P5 emphasizes the importance of support for identifying sensitive information related to financial transactions and health issues: "*there should be some support for [identifying] like financial transaction and health-related issues.*" P1 raised concerns about the system's ability to handle less directly sensitive information, questioning how the system would react to seemingly innocuous details like the names of siblings or past residences: "*I just wonder with information that isn't so directly sensitive. I wonder how the system would act. Let's say I input, I have three sisters and their names. Would that be a sort of red flag for your system? Or let's say I lived in a city when I was younger. Is that a risk because I am not living there now? So I don't know how the system would react. That's what I mean by less sensitive.*" These insights reflect participants' interest in a more comprehensive and nuanced approach to identifying and managing sensitive information.

7 Results of Case Study 2

7.1 System Usability

The evaluation of CLEAR in the context of Gmail used the same assessment method as the first case study using SUS questionnaire [2]. The results show a positive user experience, as shown in Fig 5. All participants found the system easy to use, well-integrated, and quick to learn. The confidence in using the system is high, and the information provided is useful. The policy snippets and privacy risks were appropriately matched with sensitive information.

Participants generally found the system useful for improving their awareness of privacy risks. For example, P4 said, "*So it's actually informing me about the risks. It's saying, 'If you do this, here's what happens.' So, in a way, it's guiding me through the process, showing me what I can and cannot include. It helps me understand what to share and what to avoid, making it easier to protect my privacy.*" P13 also mentioned, "*This is really good, especially from the user's perspective, because many people aren't aware of the risks. They just reply to emails, sometimes including sensitive information like their Social Security number or phone number, without thinking about the consequences. That's why it's important to clearly show what could happen if they share this information.*"

It is worth noting that participants also expressed trust in CLEAR's ability to protect their privacy, noting that although CLEAR is an LLM-based application, it identifies sensitive information locally on the user's device. Only the types of sensitive information are transmitted to the LLM, not the actual data. As P1 noted, *"[CLEAR] feels more trustworthy because it uses a solution to protect user data by not sharing the exact information with the LLM."* We refer to this as the "duality of AI", that is, AI-powered applications present privacy risks but at the same time, introduce opportunities to address privacy issues more broadly. We will expand on this point in the discussion section.

7.2 Impact on Participants' Understandings and Actions Regarding Privacy in LLMs

Similar to the findings from Case Study 1, Case Study 2 also suggests that using CLEAR increases participants' awareness of data practices and privacy risks in LLMs and leads to intentions for behavior changes. For example, some participants previously believed that the Gemini plugin processed data fully automatically, but after seeing CLEAR's notifications, they learned that human reviewers might also access their sensitive data. For example, P12 said, *"I did not know there would be any humans reviewing the data. I thought it was just an automated process."* Other participants also became aware of privacy risks related to LLMs that they were previously unaware of, such as memorization and inference risks. This increased awareness led participants to be more cautious and deliberate in protecting their private data, as we further detail in the following section.

7.3 Participants' Emotional and Behavioral Reactions to the Disclosure of Sensitive Information

7.3.1 Participants felt regretful about their past data-sharing behaviors due to the privacy risks. Participants (P6, P12, P14) were shocked and surprised by the extent of privacy risks they were previously unaware of. P14 was very surprised to see the risks and emphasized not to share personal information. P6 was not only surprised by these risks but also regretted oversharing client data with AI. She mentioned that she had to discuss this issue with her supervisor: *"Because I've already disclosed way too much information about my clients. Now I am worried about our future clients. So I'm gonna have to have a talk with my supervisor."* She also suggested an increasing feeling of empowerment from using CLEAR, stating, *"I've been really afraid of AI for a long time, but now I'm feeling hopeful that more things like this are going to come about."* These strong emotional reactions highlight the positive impact of CLEAR in raising participants' privacy awareness and encouraging changes in their privacy behaviors.

7.3.2 Participants become more cautious and deliberate in protecting their private data. Participants (P2, P3, P4, P5, P8, P9, P10, P11, P12, P14, P15) indicated their increased intention to proactively remove sensitive information from their prompts after seeing CLEAR's notifications. For instance, P2 said *"I would say I have not really paid attention to the privacy concerns in detail. But this (content in the pop-up) is actually enlightening...I think it would prompt me to keep my private information more secure."* P6 also stated, *"So I would definitely take these (sensitive information) out before asking AI to summarize that (email) for me."*

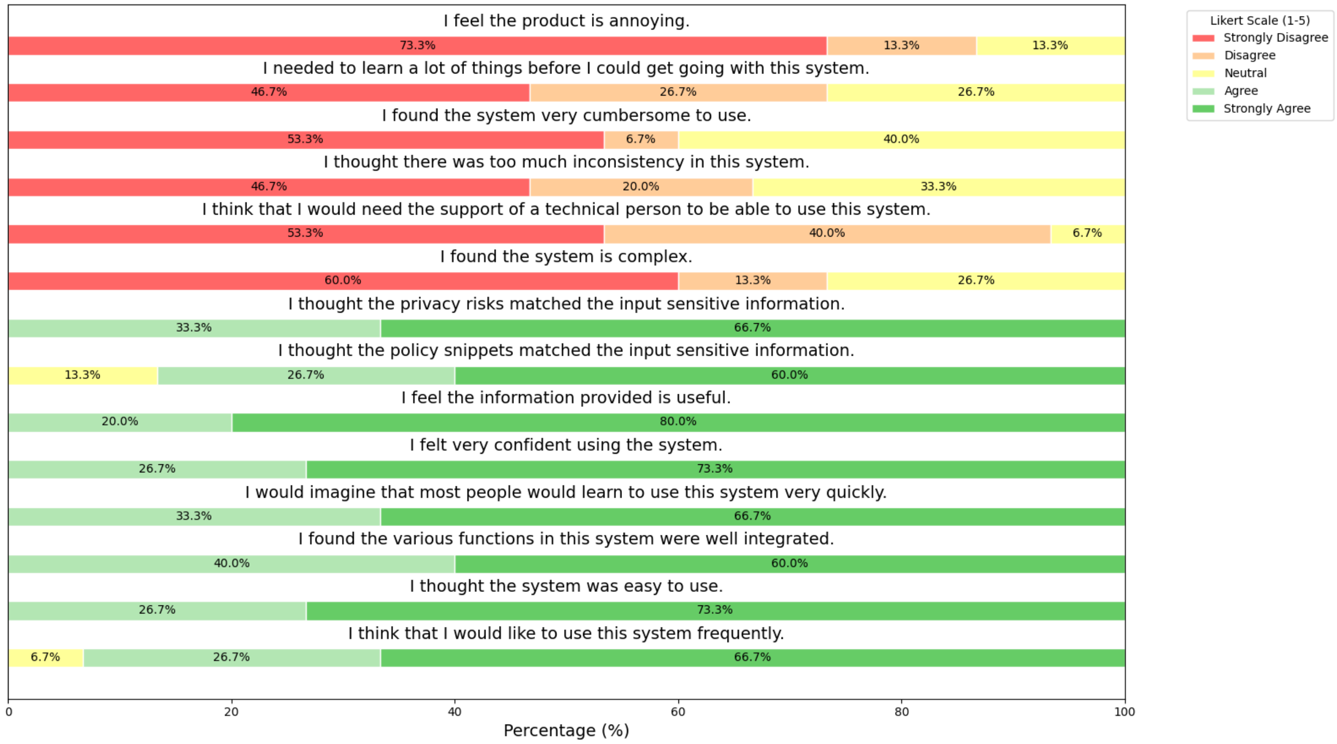


Figure 5: Result of Post-study Questionnaire for the Case Study 2.

It should be noted that unlike ChatGPT in Case Study 1 which could only access sensitive information from participants' prompts, the Gemini plugin can also retrieve sensitive information directly from emails. Therefore, removing sensitive information from the prompt will not prevent the plugin from accessing it in the emails. Participants mentioned that they would be more cautious and try to avoid sharing sensitive information with the Gemini plugin, or even not using Gemini for emails at all, as P9 mentioned, *"Maybe if I have some information I don't want to be exposed, I might not use the Gemini app."*.

Additionally, even though CLEAR focuses on providing privacy awareness in LLM-based applications, participants expressed a desire for the applications to explicitly obtain user consent before accessing their data. As P7 noted, *"It should clearly ask for permission: 'Do you allow access to your location?' with options to deny, allow, or decide later. If I decide to allow access to my location, address, IP, or work information, that should be my choice. Ultimately, nothing should happen without our consent."* To some extent, this desire indicates participants' expectation of more explicit privacy choice, which is largely lacking in today's LLM-based applications [60]. We will further discuss this point in the Discussion section.

8 Discussion

8.1 The Duality of AI in Privacy

In situating our findings within the existing literature on AI and privacy, we identified a notable tension regarding the role of AI in the management of user privacy. AI exhibits a dual nature in this context. On the one hand, recent studies indicate that AI technologies, such as LLMs, present significant privacy risks. For example, Lee et al. [27] identified twelve specific risks associated with AI, such as threats in data acquisition, data processing, data dissemination, and facilitation of cyber invasions. These risks underscore AI's potential to compromise user privacy through various channels. Moreover, the lack of transparency surrounding AI operations and its data usage contributes to diverse and often misguided mental models held by users [60]. Despite these concerns, our work revealed the opportunities and user-expressed desires to utilize AI for privacy protection. The evaluation of CLEAR provided empirical evidence supporting the positive impact of an AI-based system on enhancing users' privacy awareness and protective behaviors.

This paradox illustrates the "duality of AI" in the realm of privacy. While AI technologies inherently present privacy challenges, they simultaneously offer opportunities to assist users in managing their privacy [7, 23]. Interestingly, our findings indicated that even when participants were made aware of the potential risks associated with LLMs, they still sought AI's assistance in tasks such as "*detecting where their information actually goes*" (G3), "*scaning whatever it is that I'm typing for, specifically for privacy*" (G5), and "*tracking potential risk and take action against it*" (G1).

This phenomenon may stem from participants' perception of AI as a powerful, personalized privacy protection tool capable of delivering adaptive and contextual information based on their activities. This perspective aligns with the concept of a "personalized privacy assistant (PPA)" in the context of smartphones [1, 33] and the Internet of Things [11, 11]. However, unlike traditional PPAs that provide personalized recommendations based on privacy profiles, such as smartphone permissions [33], an AI-based system like CLEAR offers more comprehensive insights into data practices, privacy risks, and the potential consequences of users' ongoing tasks. By focusing on tasks rather than personal profiles, CLEAR adopts a less intrusive approach, reinforcing AI's role in supporting users' privacy management.

CLEAR itself capitalizes on the duality of AI. It operates under the assumption that typical LLMs, such as ChatGPT, lack transparency regarding their data practices. Users often have incomplete or varied mental models of how LLMs function [60], leading to diverse privacy concerns. The current iteration of CLEAR acts as a transparency tool, informing users about hidden data practices, associated privacy risks, and potential consequences. Our study results in Section 7.1 show that users trust CLEAR because it is more transparent and less invasive of their privacy when using LLMs. This duality allows CLEAR to effectively utilize AI to enhance transparency in AI systems.

This distinctive characteristic of AI within the privacy domain provides CLEAR with significant advantages over other approaches documented in the literature. For example, the IoT privacy nutrition labels provide data practice information to users in an easy-to-understand format [17] but would require users to proactively check the labels before purchasing the devices. Privacy Q&A uses NLP techniques to extract information from privacy policies, allowing users to ask questions about their privacy [44]. Yet, it does not provide users with the ability to associate the extracted information with the contexts of their tasks. Leveraging the duality of AI, CLEAR not only overcomes these hurdles but also leaves room to expand itself to other task domains (e.g., customer service, online medical consultant, etc.) and platforms (e.g., mobile devices, the Internet of Things).

8.2 Design Implications

Based on our findings, we propose design implications for the development of future techniques and policies aimed at enhancing users' privacy awareness related to LLM-enabled end-user applications. While in this work, we present CLEAR as an add-on to existing LLM-enabled end-user applications; we hope that their designers can consider incorporating them into the applications themselves.

8.2.1 Integrate privacy notification seamlessly with the user experience by providing contextual and real-time privacy information. The previous study [3] has demonstrated that users' privacy awareness increases significantly when privacy information is presented within the relevant context. Our study confirms this, revealing that users expect to be informed of potential privacy risks and their implications to the current task context without disruptions to their workflow. We suggest that future LLM-based systems should provide users with contextual privacy information to increase system transparency and help users make informed decisions. For instance, when users input sensitive personal information into a conversational agent, a brief, context-specific description of potential privacy risks should be displayed directly within the chat interface. Future work should also consider integrating this feature natively in conversational agents and aligning it with its specific data practices to increase its accuracy, precision, and user trust.

8.2.2 Offer users fine-grained controls over privacy information contents. Participants expressed varying preferences regarding the level of detail they want in privacy information. Some preferred concise, essential information at a glance, while others expected access to detailed explanations and full privacy policies. We recommend that future tools allow users to customize the amount of information they receive. A layered approach to presenting information, where users can define the depth of detail they want, could be especially effective in addressing individual needs and constraints.

8.2.3 Allow users to define the scope of sensitive data. Different users have varying interpretations of sensitive information. As discussed in Section 6.3.2, participants were unclear about whether certain types of "weak" sensitive information (e.g., others' personal information, past information, etc.) could be shared with LLMs. Furthermore, what is considered private often depends on the context and individual values. Therefore, it is important not only to inform users about which data is sensitive but also to give them the ability to specify the types of information they are particularly concerned about.

8.2.4 Incorporate explicit user consent mechanisms. While CLEAR focuses on increasing users' privacy awareness, participants expressed the need for explicit consent before LLMs access sensitive data. Existing LLM-based applications only obtain users' consent upon signing up for the services without the ability to change their consent during the interaction. This may become a critical issue as users' content may change when they become more aware of the potential privacy risks using systems like CLEAR. We recommend integrating these consent features into LLMs to provide fine-grained, transparent, and user-driven control, encouraging active data-sharing decisions and building trust in LLM-based applications.

8.3 Limitation and Future Work

We summarize the limitations of our work in both the study design and the system design, and we propose future steps to address these limitations.

8.3.1 Study design. Although the results of the evaluation study suggest the high usability and usefulness of our system, they are limited to a short-term study in a lab setting. A future longitudinal deployment study is needed to validate these findings in long-term and real-world settings. Our study is an initial effort to enable users to investigate relevant privacy policies and potential privacy risks when using LLMs. In the future, we will recruit a larger and more representative participant group through online recruitment platforms (e.g., Prolific), ensuring diversity across demographics, levels of technical proficiency, and privacy awareness backgrounds. Participants will install the CLEAR extension and interact with LLM conversational interfaces or LLM-based add-ons in the real context of use for 2 weeks. We plan to conduct pre-tests and post-tests with users to assess their acquisition of privacy knowledge and collect additional behavioral data when users interact with CLEAR in the deployment (similar to [48, 49]). This future longitudinal study will allow us to evaluate the ecological validity of CLEAR and the impact of its use on the user's real privacy behaviors.

8.3.2 System design. Our current CLEAR prototype serves primarily as a proof of concept to facilitate the experiment discussed in this paper. At present, it is compatible only with ChatGPT and the Gemini plugin for Gmail in order to demonstrate the feasibility of CLEAR in two primary use cases (standalone LLM conversational interfaces and the LLM add-ons for GUI applications). While ChatGPT and Gmail are the most popular applications in their category, with ChatGPT having a 60% market share [55] and Gmail holding a 53% share of the US email market[12], there are other popular LLMs (e.g., Claude) and applications (e.g., Microsoft Copilot) that the current CLEAR tool does not support. However, the key features of CLEAR do not rely on any specific technicalities of ChatGPT or Gemini for Gmail, and it is straightforward to extend CLEAR to support other chat-based LLM interfaces (e.g., Claude, web UIs for LLaMA and Qwen) and web-based LLM add-ons. To expand its scalability, we open-sourced CLEAR [13]. CLEAR's extensible architecture will make it easier to integrate CLEAR with other popular LLMs' GUI in the future.

Our system also has limited capabilities in identifying additional modalities of sensitive information (e.g., through voice interaction) and handling sensitive information contained in files such as images and PDFs uploaded to LLMs. Moreover, while we effectively use the Microsoft Presidio library to detect emails, contacts, and locations, CLEAR currently lacks the capability to promptly identify other types of personally identifiable information, such as financial and health data. To enhance CLEAR's functionality, we plan to investigate using small-scale, open-source language models to identify sensitive information on the edge. Recent advancements [29, 61] allow these models to run locally on the client side, allowing for more accurate detection of sensitive information across a broader range of types, while keeping user data on-device and mitigating privacy risks.

Currently, CLEAR uses text to summarize privacy policies and generate privacy risks as an initial effort to enable users to investigate relevant privacy policies and potential risks when using LLMs. For users with limited privacy literacy or technical expertise, textual descriptions may not be the most efficient way to communicate the implications of privacy practices or the significance of identified risks. With the development of multimodal LLMs, there is an opportunity to overcome these limitations by incorporating more interactive and accessible explanations. Future work could focus on integrating multimodal features into CLEAR, such as interactive tutorials or visual summaries that highlight key aspects of privacy policies and risks. For example, users could explore how their data might be processed or shared through animations or flowcharts that depict data flows. Integrating multimodal LLMs into CLEAR has the potential to transform how privacy policies and risks are communicated, making them more transparent, engaging, and comprehensible for a wider audience.

9 Conclusion

In response to the emerging privacy risks posed by the use of LLMs in end-user applications, we conducted five co-design workshops to study user needs, preferences, and constraints relevant to making informed privacy decisions when interacting with LLM-enabled end-user applications. Based on our findings, we developed CLEAR, a tool aimed at enhancing user understanding of privacy policies and associated risks in LLM-enabled applications. We conducted two user studies, one with the representative standalone LLM application ChatGPT and another with the Gemini ad-don for Gemail, demonstrating that CLEAR significantly improves privacy awareness and encourages safer data-sharing practices. The insights derived from our research offer important design implications for the development of AI-based tools that empower users to manage their privacy more effectively.

Acknowledgments

This work was supported in part by a Google Research Scholar Award, a Notre Dame-IBM Technology Ethics Lab Award, a Google PSS Privacy Research Award, NSF Grant CNS-2426395, NSF Grant CNS-2426397, and NSF Grant CNS-2232653. C. Chen and Y. Ye's work was partially supported by the NSF under grants IIS-2321504, IIS-2334193, IIS-2217239, CNS-2426514, CNS-2203261, and CMMI-2146076. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors. We would like to thank Allen Yilun Lin for useful discussions.

References

[1] Hazim Almuhimedi, Florian Schaub, Norman Sadeh, Idris Adjerid, Alessandro Acquisti, Joshua Gluck, Lorrie Faith Cranor, and Yuvraj Agarwal. 2015. Your location has been shared 5,398 times! A field study on mobile app privacy nudging. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems.* 787–796.

[2] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies* 4, 3 (2009), 114–123.

[3] Mike Bergmann. 2008. Testing privacy awareness. In *IFIP Summer School on the Future of Identity in the Information Society.* Springer, 237–253.

[4] Meryl Brod, Laura E Tesler, and Torsten L Christensen. 2009. Qualitative research and content validity: developing best practices based on science and experience. *Quality of life research* 18 (2009), 1263–1278.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[6] 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).* 2633–2650.

[7] Chaoran Chen, Weijun Li, Wenxin Song, Yanfang Ye, Yaxing Yao, and Toby Jia-Jun Li. 2024. An Empathy-Based Sandbox Approach to Bridge the Privacy Gap among Attitudes, Goals, Knowledge, and Behaviors. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.* 1–28.

[8] Cameron Coles. 2023. *4.2% of Workers Have Pasted Company Data into ChatGPT.* https://www.cyberhaven.com/blog/4-2-of-workers-have-pasted-company-data-into-chatgpt Accessed: 2024-05-26.

[9] Tianyu Cui, Yanling Wang, Chuanpu Fu, Yong Xiao, Sijia Li, Xinhao Deng, Yunpeng Liu, Qinglin Zhang, Ziyi Qiu, Peiyang Li, et al. 2024. Risk taxonomy, mitigation, and assessment benchmarks of large language model systems. *arXiv preprint arXiv:2401.05778* (2024).

[10] Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023. LLM-in-the-loop: Leveraging large language model for thematic analysis. *arXiv preprint arXiv:2310.15100* (2023).

[11] Anupam Das, Martin Degeling, Daniel Smullen, and Norman Sadeh. 2018. Personalized privacy assistants for the internet of things: Providing users with notice

[12] https://techreport.com/statistics/software-web/gmail-statistics/

[13] https://github.com/CRChenND/CLEAR

and choice. *IEEE Pervasive Computing* 17, 3 (2018), 35–46.

[12] Badhan Chandra Das, M Hadi Amini, and Yanzhao Wu. 2024. Security and privacy challenges of large language models: A survey. *arXiv preprint arXiv:2402.00888* (2024).

[13] 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* (2023).

[14] CM Downey, Wei Dai, Huseyin A Inan, Kim Laine, Saurabh Naik, and Tomasz Religa. 2022. Planting and mitigating memorized content in Predictive-Text language models. *arXiv preprint arXiv:2212.08619* (2022).

[15] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2016. Calibrating noise to sensitivity in private data analysis. *Journal of Privacy and Confidentiality* 7, 3 (2016), 17–51.

[16] Kennedy Edemacu and Xintao Wu. 2024. Privacy preserving prompt engineering: A survey. *arXiv preprint arXiv:2404.06001* (2024).

[17] Pardis Emami-Naeini, Janarth Dheenadhayalan, Yuvraj Agarwal, and Lorrie Faith Cranor. 2021. An informative security and privacy "nutrition" label for internet of things devices. *IEEE Security & Privacy* 20, 2 (2021), 31–39.

[18] Denis Feth. 2017. Transparency through Contextual Privacy Statements. In *Veröffentlicht durch die Gesellschaft für Informatik*. https://doi.org/10.18420/muc2017-ws05-0406

[19] Ece Gumusel, Kyrie Zhixuan Zhou, and Madelyn Rose Sanfilippo. 2024. User Privacy Harms and Risks in Conversational AI: A Proposed Framework. *arXiv preprint arXiv:2402.09716* (2024).

[20] Huseyin A Inan, Osman Ramadan, Lukas Wutschitz, Daniel Jones, Victor Rühle, James Withers, and Robert Sim. 2021. Training data leakage analysis in language models. *arXiv preprint arXiv:2101.05405* (2021).

[21] Shotaro Ishihara. 2023. Training data extraction from pre-trained language models: A survey. *arXiv preprint arXiv:2305.16157* (2023).

[22] Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2022. Knowledge unlearning for mitigating privacy risks in language models. *arXiv preprint arXiv:2210.01504* (2022).

[23] Shivani Kapania, Ruiyi Wang, Toby Jia-Jun Li, Tianshi Li, and Hong Shen. 2024. " I'm categorizing LLM as a productivity tool": Examining ethics of LLM use in HCI research practices. *arXiv preprint arXiv:2403.19876* (2024).

[24] Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2024. Propile: Probing privacy leakage in large language models. *Advances in Neural Information Processing Systems* 36 (2024).

[25] Youjeong Kim and S Shyam Sundar. 2012. Anthropomorphism of computers: Is it mindful or mindless? *Computers in Human Behavior* 28, 1 (2012), 241–250.

[26] Nir Kshetri. 2023. Cybercrime and privacy threats of large language models. *IT Professional* 25, 3 (2023), 9–13.

[27] Hao-Ping Lee, Yu-Ju Yang, Thomas Serban Von Davier, Jodi Forlizzi, and Sauvik Das. 2024. Deepfakes, Phrenology, Surveillance, and More! A Taxonomy of AI Privacy Risks. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–19.

[28] Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, Jie Huang, Fanpu Meng, and Yangqiu Song. 2023. Multi-step jailbreaking privacy attacks on chatgpt. *arXiv preprint arXiv:2304.05197* (2023).

[29] Qiyu Li, Jinhe Wen, and Haojian Jin. 2024. Governing Open Vocabulary Data Leaks Using an Edge LLM through Programming by Example. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 4 (2024), 1–31.

[30] Tianshi Li, Sauvik Das, Hao-Ping Lee, Dakuo Wang, Bingsheng Yao, and Zhiping Zhang. 2024. Human-Centered Privacy Research in the Age of Large Language Models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–4.

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

[32] 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)*. 4188–4203.

[33] Bin Liu, Mads Schaarup Andersen, Florian Schaub, Hazim Almuhimedi, Shikun Aerin Zhang, Norman Sadeh, Yuvraj Agarwal, and Alessandro Acquisti. 2016. Follow my recommendations: A personalized privacy assistant for mobile app permissions. In *Twelfth symposium on usable privacy and security (SOUPS 2016)*. 27–41.

[34] Yuwen Lu, Chao Zhang, Yuewen Yang, Yaxing Yao, and Toby Jia-Jun Li. 2024. From Awareness to Action: Exploring End-User Empowerment Interventions for Dark Patterns in UX. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 59 (April 2024), 41 pages. https://doi.org/10.1145/3637336

[35] Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. 2023. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 346–363.

[36] Pingchuan Ma, Rui Ding, Shuai Wang, Shi Han, and Dongmei Zhang. 2023. InsightPilot: An LLM-empowered automated data exploration system. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 346–352.

[37] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–23.

[38] Helen Nissenbaum. 2004. Privacy as contextual integrity. *Wash. L. Rev.* 79 (2004), 119.

[39] Anna-Marie Ortloff, Maximiliane Windl, Valentin Schwind, and Niels Henze. 2020. Implementation and in situ assessment of contextual privacy policies. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 1765–1778.

[40] Shidong Pan, Zhen Tao, Thong Hoang, Dawen Zhang, Tianshi Li, Zhenchang Xing, Sherry Xu, Mark Staples, Thierry Rakotoarivelo, and David Lo. 2024. {A New Hope}: Contextual Privacy Policies for Mobile Applications and An Approach Toward Automated Generation. *arXiv preprint arXiv:2402.14544* (2024).

[41] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286* (2022).

[42] Natalia Ponomareva, Hussein Hazimeh, Alex Kurakin, Zheng Xu, Carson Denison, H Brendan McMahan, Sergei Vassilvitskii, Steve Chien, and Abhradeep Guha Thakurta. 2023. How to dp-fy ml: A practical guide to machine learning with differential privacy. *Journal of Artificial Intelligence Research* 77 (2023), 1113–1201.

[43] Lorenzo Porcelli, Massimo Ficco, and Francesco Palmieri. 2023. Mitigating User Exposure to Dark Patterns in Cookie Banners Through Automated Consent. In *International Conference on Computational Science and Its Applications*. Springer, 145–159.

[44] Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining computational and legal perspectives. *arXiv preprint arXiv:1911.00841* (2019).

[45] Maximilian Schmidhuber and Udo Kruschwitz. 2024. Llm-based synthetic datasets: Applications and limitations in toxicity detection. *LREC-COLING 2024* (2024), 37.

[46] Victoria Smith, Ali Shahin Shamsabadi, Carolyn Ashurst, and Adrian Weller. 2023. Identifying and mitigating privacy risks stemming from language models: A survey. *arXiv preprint arXiv:2310.01424* (2023).

[47] Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2023. Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298* (2023).

[48] Peter Story, Daniel Smullen, Rex Chen, Yaxing Yao, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. 2022. Increasing adoption of tor browser using informational and planning nudges. *Proceedings on Privacy Enhancing Technologies* (2022).

[49] Peter Story, Daniel Smullen, Yaxing Yao, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. 2021. Awareness, adoption, and misconceptions of web privacy tools. *Proceedings on Privacy Enhancing Technologies* (2021).

[50] Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023. Can ChatGPT replace traditional KBQA models? An in-depth analysis of the question answering performance of the GPT LLM family. In *International Semantic Web Conference*. Springer, 348–367.

[51] Thomas Vakili, Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. 2022. Downstream task performance of bert models pre-trained using automatically de-identified clinical data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 4245–4252.

[52] Froukje Sleeswijk Visser. 2005. Contextmapping: experiences from practice. *CoDesign* (2005).

[53] Xingyi Wang, Xiaozheng Wang, Sunyup Park, and Yaxing Yao. 2025. Mental Models of Generative AI Chatbot Ecosystems. *30th International Conference on Intelligent User Interfaces (IUI'25* (2025).

[54] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[55] Chris Westfall. 2023. New Research Shows ChatGPT Reigns Supreme in AI Tool Sector. *Forbes* (2023). https://www.forbes.com/sites/chriswestfall/2023/11/16/new-research-shows-chatgpt-reigns-supreme-in-ai-tool-sector/?sh=52ff59c050e9

[56] Maximiliane Windl, Niels Henze, Albrecht Schmidt, and Sebastian S Feger. 2022. Automating contextual privacy policies: Design and evaluation of a production tool for digital consumer privacy awareness. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.

[57] Yaxing Yao, Justin Reed Basdeo, Smirity Kaushik, and Yang Wang. 2019. Defending my castle: A co-design study of privacy mechanisms for smart homes. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[58] Yaxing Yao, Justin Reed Basdeo, Oriana Rosata Mcdonough, and Yang Wang. 2019. Privacy perceptions and designs of bystanders in smart homes. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[59] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing* (2024), 100211.

[60] Zhiping Zhang, Michelle Jia, Hao-Ping Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. 2024. "It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–26.

[61] Jijie Zhou, Eryue Xu, Yaoyao Wu, and Tianshi Li. 2024. Rescriber: Smaller-LLM-Powered User-Led Data Minimization for Navigating Privacy Trade-offs in LLM-Based Conversational Agent. *arXiv preprint arXiv:2410.11876* (2024).

[62] Jakub Złotowski, Diane Proudfoot, Kumar Yogeeswaran, and Christoph Bartneck. 2015. Anthropomorphism: opportunities and challenges in human–robot interaction. *International journal of social robotics* 7 (2015), 347–360.

A. Codebook for the participatory design [1]

(1) User needs
  (a) Improve privacy Literacy
    (i) Force users to spend more time understanding policies (3)
    (ii) Show connection between data usage and system features (3)
    (iii) Have a tutorial to go through all security features in the system (2)
    (iv) Raise users' awareness by showing the value of their data (2)
    (v) Simulating the scams to help people learn to react to them (2)
    (vi) Have courses to improve people's privacy literacy (1)
  (b) Use AI to protect privacy
    (i) Inspect sensitive info in user input (4)
    (ii) Have LLMs unlearn data (1)
    (iii) Use AI to monitor users and defend against hacking (10)
    (iv) Detect data traffic and block risky ones (1)
  (c) Enable users to take action
    (i) Help users do data sanitation (1)
    (ii) Help users recognize and manage the data they share (10)
    (iii) Provide users with an option to either skip reading it or read it (2)
    (iv) Give users options to accept, deny, or modify info to submit (4)
  (d) Information presentation
    (i) Show reasons and consequences of privacy data usage
      A. Use animation to show the consequences of giving consent (2)
      B. Show consequences of sharing certain private data (5)
    (ii) Information should be adjusted to user groups
      A. Explain the privacy segments for people in every generation (1)
      B. The tool should be accessible to all user groups (7)
    (iii) Information should be compact and easy to understand
      A. Add tooltips to show definitions of words in policy (1)
      B. Visualize important info on one page and show others in links (5)
      C. Add link to the detailed information (6)
      D. Use simple language to summarize the information (8)
      E. Show information concisely (3)
  (e) Contextual privacy policy and risk alert
    A. Summarize the share of private data and potential privacy risks (1)
    B. Indicate users of sensitive information and possible reactions (5)
    C. Notify relevant policy segments and consequences (6)
    D. Alert themselves if a privacy breach happens (8)
    E. Show brief and in-context privacy notice (2)

[1] The numbers in parentheses indicate the number of participants associated with each code in the codebook.

(2) Existing privacy protection methods
  (a) Self-data sanitation
    (i) Disguise real location and online behavior data with fake ones (1)
    (ii) Use fake name online (2)
    (iii) Anonymize data and disassociate data (3)
  (b) Self-privacy education
    (i) Check app permission regularly (2)
    (ii) Educate themselves on privacy risks (5)
  (c) Secure browsing practices
    (i) Use different usernames/passwords for every site (8)
    (ii) More strict authentication (4)
    (iii) Use a more secure browser to search (1)
  (d) Limit data sharing
    (i) Have control over the data being shared (8)
    (ii) Store privacy data locally (1)
    (iii) Not sharing sensitive information to risky app or people (11)
    (iv) Limit the data types that users can upload to the system (2)

(3) Things in privacy policies that people care about
  (a) Show detailed information about who will use the privacy data (6)
  (b) How the data access influences users
    (i) Influence of privacy policy updates (2)
    (ii) Security and privacy risks (2)
  (c) What data is being accessed
    (i) What kind of data the app will access (4)
    (ii) Payment and transaction (1)
    (iii) Access to contact information (1)
    (iv) Frequently used service (1)
    (v) Tracking of online behavior (2)
  (d) Why is data being accessed (3)

(4) Reasons for not reading privacy policies
  (a) Privacy resignation
    (i) Users think service providers don't want them to read (1)
    (ii) Users think their data will always be leaked (8)
    (iii) Users have to use the service (2)
    (iv) Users think all privacy policies are the same (5)
  (b) Privacy policy is difficult to understand
    (i) Design issue/not user friendly (2)
    (ii) Difficult to understand (6)
    (iii) Users feel uncertain to identify (2)
  (c) Privacy policy is too long
    (i) Privacy policies are too long (10)
    (ii) Users don't want to spend a lot of time reading and understanding (2)

B. Structured Privacy Knowledge

(1) **Privacy Risk:** Data breaches and sale of personal data
- **Cause:** LLM-based CAs operating on the cloud
- **Influence:** Users losing control over their chat logs
- **Evidence:** Most popular LLM-based CAs operate on the cloud

(2) **Privacy Risk:** Memorization risks
- **Cause:** LLMs memorizing details in the training data
- **Influence:** Sensitive information being leaked in response to prompts
- **Evidence:** LLMs using user data to train their models periodically

(3) **Privacy Risk:** Disclosure of personally identifiable information (PII)
- **Cause:** Users disclosing their own and others' data in conversations
- **Influence:** Implicating interdependent privacy issues
- **Evidence:** Users disclosing various types of PII in chat histories

(4) **Privacy Risk:** Dark patterns in opt-out interfaces
- **Cause:** Discouraging users from exercising privacy controls
- **Influence:** Users feeling they have to sacrifice privacy for benefits
- **Evidence:** Opt-out interfaces linking privacy and utility loss

(5) **Privacy Risk:** Extraction of personal attributes from text
- **Cause:** LLMs lack commonsense about social privacy norms
- **Influence:** Malicious actors can infer personal attributes from seemingly harmless text
- **Evidence:** Inference of location based on text mentioning a specific traffic maneuver

(6) **Privacy Risk:** Lack of understanding of privacy norms
- **Cause:** LLMs lack commonsense about social privacy norms
- **Influence:** Difficulty in keeping secrets and protecting privacy
- **Evidence:** Models can be easily tricked by third-party adversaries to ignore privacy-protecting instructions

(7) **Privacy Risk:** User sharing sensitive information with LLM-based CAs
- **Cause:** High utility and human-like interactions of LLM-based CAs
- **Influence:** Users sharing sensitive and personally identifiable information
- **Evidence:** Users constantly facing challenges in protecting their privacy due to flawed mental models and dark patterns in privacy management features

(8) **Privacy Risk:** Exposure risk
- **Cause:** AI technologies can generate human-like media, such as deepfake pornography, without consent
- **Influence:** AI technologies can lead to the unauthorized dissemination of sensitive or private information
- **Evidence:** Twitch streamer QTCinderella's plea to stop spreading links to AI-generated deepfake pornography

(9) **Privacy Risk:** Phrenology/physiognomy risk
- **Cause:** AI can learn arbitrary classification functions and potentially infer sensitive attributes like sexual orientation from physical features
- **Influence:** AI technologies can perpetuate harmful stereotypes and discrimination based on physical appearance
- **Evidence:** The belief that AI can automatically detect things like sexual orientation from physical attributes

(10) **Privacy Risk:** Surveillance risk
- **Cause:** Facial recognition classifiers require large amounts of face data, leading to uncritical data collection practices
- **Influence:** AI technologies can enable widespread surveillance and tracking of individuals without their consent
- **Evidence:** Collection of face scans in airports for facial recognition purposes

C. Codebooks [1]

Case Study 1

(1) User feedback
- (a) Identify the access to specific services, such as text messaging or conversations (1)
- (b) Provide users options to look into more details (1)
- (c) Show potential solutions to users (2)
- (d) Identify more types of sensitive info (image/PDF/financial/health) (4)
- (e) Why the number of policy snippets differ (1)
- (f) How to definite the scope of sensitive information (3)

(2) Changes brought yo users after using the tool
- (a) Compare benefit and risk, if the risk is not bad, accept it (5)
- (b) Intention for behavior change (2)
- (c) Raise more privacy concerns (2)
- (d) Be aware of new reasons for data accessing (2)
- (e) Be aware of new privacy risks (4)
- (f) Clarify privacy risks (3)
- (g) Be aware of new data accessors (11)

(3) Reaction to the prompt after seeing the tool
- (a) Use fake/dummy info to replace sensitive info (7)
- (b) Use placeholder to replace sensitive info (1)
- (c) Remove sensitive info (5)

Case Study 2

(1) Participants' emotion
- (a) Feels hopeful and excited about the extension (1)
- (b) Be shocked and surprised about the privacy risks (5)
- (c) Regrets oversharing information with AI (1)

(2) Behavior change
- (a) Refuse to use LLM if any sensitive information is involved (1)
- (b) Remain no change due to privacy resignation (2)
- (c) Desire to keep certain sensitive information private (4)
- (d) Actively removing personal data from their prompts (8)
- (e) Be more cautious and deliberate in protecting their privacy data (14)

(3) Participant's suggestions
- (a) Shows consequences of privacy data leak(2)
- (b) The information can be more concise (4)
- (c) Identify sensitive information in the attachments(1)
- (d) Ask for the client's consent about sharing the sensitive info (1)
- (e) Useful to know before sending an email out (1)
- (f) The extension should allow users to not share personal info (1)
- (g) Websites should get user consent before accessing personal data (1)
- (h) Participants want to know more about how data may be leaked (3)
- (i) Add 2FA or passkeys before sharing personal info (1)

(4) Comment on usability and usefulness
- (a) Useful to highlight sensitive info, especially for long emails (1)
- (b) Help verify privacy hypothesis (1)
- (c) Plans to inform their supervisor about the data leakage (1)
- (d) Highly impressed and appreciative of the detailed information (1)
- (e) The extension is easy to use and the function is positive (15)
- (f) The extension is trustful since it protects users' exact data (2)
- (g) The extension is explanatory and easy to understand (8)

(5) Improve participants' privacy awareness
- (a) Improve users' awareness about privacy risks (2)
- (b) Sensitive info may be collected without explicit consent (2)
- (c) Understand flawed mental model and dark pattern (1)
- (d) Sensitive information can be used for controlled access (2)
- (e) LLMs may lack conformity to social privacy norms (2)
- (f) Someone can prompt and LLM to reveal personal info (3)
- (g) Memorization risk (10)
- (h) The reason why Google want to access phone number or other information (2)
- (i) Human reviewers might access sensitive data (7)

[1] The numbers in parentheses indicate the number of participants associated with each code in the codebook.