



# Rescriber: Smaller-LLM-Powered User-Led Data Minimization for LLM-Based Chatbots

Jijie Zhou  
Khoury College of Computer Sciences  
Northeastern University  
Boston, Massachusetts, USA  
seazon96zhou@gmail.com

Yaoyao Wu  
Khoury College of Computer Sciences  
Northeastern University  
Boston, Massachusetts, USA  
wuyaoyao0309@gmail.com

Eryue Xu  
Khoury College of Computer Sciences  
Northeastern University  
Boston, Massachusetts, USA  
er.xu@northeastern.edu

Tianshi Li  
Khoury College of Computer Sciences  
Northeastern University  
Boston, Massachusetts, USA  
tia.li@northeastern.edu



# Rescriber: 基于小型LLM的用户主导数据最小化方案 基于LLM的聊天机器人

周继杰 Khoury计算机科学学院  
院 东北大学 美国马萨诸塞州波士顿 seazon96zhou@gmail.com

吴瑶瑶 Khoury计算机科学学院  
院 东北大学 美国马萨诸塞州波士顿 wuyaoyao0309@gmail.com

徐尔悦 Khoury计算机科学学院  
东北大学 美国马萨诸塞州波士顿 er.xu@northeastern.edu

李天石 Khoury计算机科学学院  
院 东北大学 美国马萨诸塞州波士顿 tia.li@northeastern.edu

## Abstract

The proliferation of LLM-based conversational agents has resulted in excessive disclosure of identifiable or sensitive information. However, existing technologies fail to offer perceptible control or account for users' personal preferences about privacy-utility tradeoffs due to the lack of user involvement. To bridge this gap, we designed, built, and evaluated Rescriber, a browser extension that supports user-led data minimization in LLM-based conversational agents by helping users detect and sanitize personal information in their prompts. Our studies (N=Rescriber) showed that Rescriber helped users reduce unnecessary disclosure and addressed their privacy concerns. Users' subjective perceptions of the system powered by Llama3-8B were on par with that by GPT-4o. The comprehensiveness and consistency of the detection and sanitization emerge as essential factors that affect users' trust and perceived protection. Our findings confirm the viability of smaller-LLM-powered, user-facing, on-device privacy controls, presenting a promising approach to address the privacy and trust challenges of AI.

## CCS Concepts

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

## Keywords

privacy, security, LLM, AI, chatbot, PII, ChatGPT

## ACM Reference Format:

Jijie Zhou, Eryue Xu, Yaoyao Wu, and Tianshi Li. 2025. Rescriber: Smaller-LLM-Powered User-Led Data Minimization for LLM-Based Chatbots. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan*. ACM, New York, NY, USA, 28 pages. <https://doi.org/10.1145/3706598.3713701>



This work is licensed under a Creative Commons Attribution 4.0 International License.  
*CHI '25, Yokohama, Japan*  
© 2025 Copyright held by the owner/author(s).  
ACM ISBN 979-8-4007-1394-1/25/04  
<https://doi.org/10.1145/3706598.3713701>

## 1 Introduction

The emergence of large language models like OpenAI's GPT [36], Google's Gemini [17], and Meta's Llama [48] has enabled LLM-based conversational agents that understand the context and can generate human-like responses. As products with high capacity for handling natural language tasks and user-friendly interfaces, such as ChatGPT, come into play, people increasingly rely on them for a variety of tasks in personal and professional contexts [5, 10, 57].

Consequently, a large amount of personal or sensitive disclosure has been observed in users' conversations with ChatGPT-like LLM-based conversational agents [30, 57], which leads to severe privacy threats in the event of database breaches and data extraction attacks targeting the memorization vulnerabilities of LLMs [8, 9, 33, 54]. This includes personally identifiable information (PII) like names and email addresses, as well as non-PII sensitive topics such as sexual preferences and drug use [30]. Notably, parts of the disclosure are unnecessary to the main task, like hyper-detailed information: "Help me generate a one-day itinerary in NYC, I live at 153 W 57th St, New York, NY 10019", or irrelevant information: "Please help me proofread the following email to my colleague peter ([peter.parker@spider.com](mailto:peter.parker@spider.com))".

The excessive and sensitive disclosure issues pose a challenge to the classic data privacy principle, *data minimization* [15, 37, 41]. Although researchers and practitioners have employed data sanitization [14, 35, 38] and private execution in trusted hardware [1, 49] to achieve the goal of removing PII at the training stage and avoiding unnecessary data retention at the inference stage, there are several inherent limitations of the techniques due to the lack of involvement of the end-users. First, the privacy and utility needs vary by context and person, meaning post-hoc data sanitization may not remove all the information the user found sensitive or unnecessary. Second, once data leaves the device, users lose direct control over their information. Existing techniques operate in a concealed way, making it difficult for users to form accurate mental models, reduce privacy concerns, and trust LLM-based agents with sensitive information. These issues reveal a research gap in designing user-friendly tools that enhance users' ability to navigate privacy risks and minimize unnecessary information disclosure, specifically enabling *user-led data minimization*.

## 摘要

基于LLM的对话代理的普及导致了可识别或敏感信息的过度披露。然而，由于缺乏用户参与，现有技术无法提供可感知的控制或考虑用户对隐私-效用权衡的个人偏好。为了弥合这一差距，我们设计、构建并评估了Rescriber，这是一款浏览器扩展，通过帮助用户检测和清理提示中的个人信息，支持在基于LLM的对话代理中实现用户主导的数据最小化。我们的研究（N=Rescriber）表明，Rescriber帮助用户减少了不必要的披露并解决了他们的隐私顾虑。用户对由Llama3-8B模型驱动的系统的主观感知与GPT-4o相当。检测与清理的全面性和一致性成为影响用户信任与感知保护的关键因素。

我们的研究结果证实了由小型LLM驱动的、面向用户的设备端隐私控制的可行性，为解决人工智能的隐私与信任挑战提供了一种有前景的方法。

## CCS概念

- 安全与隐私 → 隐私保护; • 以人为本的计算 → 交互式系统与工具。

## 关键词

隐私, 安全, 大型语言模型, 人工智能, 聊天机器人, 个人身份信息, ChatGPT

### 美国计算机协会参考文献格式:

周继杰、徐尔悦、吴瑶瑶和李天石。2025。Rescriber：基于小型LLM的用户主导数据最小化方案用于LLM驱动的聊天机器人。发表于计算机系统中的人为因素会议（CHI '25），2025年4月26日至5月1日，日本横滨。美国计算机协会，美国纽约州纽约市，28页。

<https://doi.org/10.1145/3706598.3713701>



本作品采用知识共享署名4.0国际许可协议进行许可。  
*CHI '25, 日本横滨 © 2025 版权归所有者/作者所有。*  
美国计算机协会国际标准书号  
979-8-4007-1394-1/25/04  
<https://doi.org/10.1145/3706598.3713701>

## 1 引言

OpenAI的GPT [36]等大型语言模型的出现谷歌的Gemini [17]，和Meta的Llama [48]已赋能基于大型语言模型的对话代理，这些代理能理解上下文并生成类人响应。随着诸如ChatGPT等具备强大自然语言处理能力和用户友好界面的产品涌现，人们在个人与专业场景[5, 10, 57]中愈发依赖它们完成各类任务。

因此，在用户与ChatGPT类大型语言模型对话代理的交互中，已观察到大量个人或敏感信息披露[30, 57]，若发生数据库泄露及针对大型语言模型记忆漏洞的数据提取攻击[8, 9, 33, 54]，将引发严重的隐私威胁。

这包括个人身份信息（PII），如姓名和电子邮件地址，以及非个人身份信息的敏感话题，如性取向和药物使用[30]。值得注意的是，部分披露内容对主要任务而言是不必要的，例如过度详细的细节信息：

"帮我生成一个纽约市的一日行程，我住在纽约州纽约市西57街153号，邮编10019"，或是无关信息："请帮我校对以下发送给同事彼得（[peter.parker@spider.com](mailto:peter.parker@spider.com)）的电子邮件"。

过度且敏感的披露问题对经典的数据隐私原则——数据最小化[15, 37, 41]构成了挑战。

尽管研究者和从业者已采用数据清理[14, 35, 38]与可信硬件[1, 49]中的私有执行技术，旨在训练阶段去除PII并在推理阶段避免不必要的数据留存，但由于缺乏终端用户的参与，这些技术存在若干固有局限。首先，隐私与效用需求因上下文和个人而异，这意味着事后数据清理可能无法完全移除用户认为敏感或不必要的信息。其次，一旦数据离开设备，用户便失去对其信息的直接控制权。现有技术以隐蔽方式运作，使用户难以建立准确的心理模型、降低隐私顾虑并信任基于LLM的代理处理敏感信息。这些问题揭示了研究空白：亟需设计用户友好工具以增强用户应对隐私风险和最小化不必要的信息披露的能力。

特别是支持用户主导的数据最小化。



**Figure 1: A snapshot of Rescriber user experience when using ChatGPT for data analysis.** Rescriber displays a tooltip that highlights the detected personal information in the user’s message (A), and offers a control panel (B) where users can either replace the information with a placeholder (B1), abstract the information to a more general version (B2), or revert the actions (B3). All the changes would be immediately effective on the message that would be sent out in the input box (C). In this example, the user replaced the detected names (e.g., James Williams) to corresponding placeholders (e.g., [NAME4]). For the sent messages (D1) and ChatGPT’s responses (D2), Rescriber replaces the placeholders with the original PII to increase readability and facilitate copying. When the user mouses over the highlighted PII, the placeholders used in the actual messages sent and received are revealed (D).

Prior research has emphasized the trade-offs among privacy, utility, and convenience as a major challenge to achieving privacy-preserving disclosure behaviors to LLM-based conversational agents [57]. This is especially true when it comes to the excessive disclosure issue. Excessive disclosure often arises when users paste texts with semantically unrelated information they didn’t notice or find too cumbersome to remove. Zhang et al. [57] found that users sometimes manually sanitize inputs, but the manual actions fall short due to forgetfulness and tediousness. They also found that users were pessimistic about preserving privacy while reaping the benefits of AI, and users were neither capable of identifying all the opportunities associated with minimizing excessive sensitive disclosures nor willing to invest in efforts and time to do that [57]. Hence, we argue that a tool that supports user-led data minimization should act in a way that *maintains response quality* and *incurs minimal overhead*.

In this work, we make an initial foray into designing user-led data minimization support for LLM-based conversational agents to tackle the aforementioned challenges. Specifically, we aim to design a browser extension that leverages AI to streamline the process of identifying and sanitizing unnecessary disclosures of personal information, as LLMs significantly outperform traditional NLP methods for detecting PII [7]. There are several key challenges in achieving this goal. First, accurate detection and sanitization recommendations are essential for effective support for user-led data minimization [55]. According to the neural scaling laws [42, 52], using larger models increases the accuracy, while it also imposes

higher computational costs and raises the bar of the user’s device for deploying the system locally. Furthermore, it’s difficult to determine the sweet spots of leveraging large language models to achieve the ultimate goal of empowering users to stay in control over their data while maintaining the response utility. Lastly, it is unclear how different approaches to data minimization affect users’ privacy perceptions and behaviors. Formally, we aim to investigate the following research questions on core efficacy, adoption intention, and impact on users:

- RQ1 How can we design a tool to support effective data minimization and reduce users’ privacy concerns?
- RQ2 How can we design a tool for user-led data minimization that users trust and intend to use?
- RQ3 How does the support for user-led data minimization affect users’ privacy awareness and behaviors?

We present Rescriber, a browser extension for Chrome, which detects and highlights potential personal information disclosures, and then provides users with options to redact or abstract these details before sending a message. Redaction replaces personal terms with entity placeholders (e.g., “Alex” becomes “[NAME1]”), while abstraction replaces them with generalized and contextually appropriate information (e.g., “my colleague”). When the address is highlighted as personal information in “Help me generate a one-day itinerary in NYC, I live at 153 W 57th St, New York, NY 10019”, we could abstract the information and the message would turn into “Help me generate a one-day itinerary in NYC, I live near Central Park”, this way, the contextual information still exists, but with a

**图1：用户使用ChatGPT进行数据分析时Rescriber的用户体验截图。** Rescriber显示一个工具提示，高亮检测到的用户消息中的个人信息（A），并提供一个控制面板（B），用户可以在其中选择用占位符替换信息（B1）、将信息抽象为更通用的版本（B2）或撤销操作（B3）。所有更改将立即生效于输入框（C）中即将发送的消息。在此示例中，用户将检测到的姓名（如詹姆斯·威廉姆斯）替换为对应的占位符（如[NAME4]）。对于已发送的消息（D1）和ChatGPT的响应（D2），Rescriber会将占位符还原为原始个人身份信息以提升可读性并便于复制。当用户悬停在高亮的个人身份信息上时，实际收发消息中使用的占位符会显示出来（D）。

先前的研究强调了隐私之间的权衡，效用和便利性被视为实现隐私保护性披露行为向基于LLM的对话代理[57]的主要挑战。这在涉及过度披露问题时尤为明显。当用户粘贴包含语义无关信息且未注意或认为过于繁琐而未能移除的文本时，常会发生过度披露。张等人[57]发现用户有时会手动清理输入，但由于遗忘和繁琐性，这些手动操作往往效果不佳。他们还发现，用户对于享受人工智能益处的同时保护隐私持悲观态度，且用户既无法识别所有与最小化过度敏感信息披露相关的机会，也不愿投入精力和时间去做[57]。

因此，我们认为支持用户主导的数据最小化的工具应以保持响应质量和产生最小开销的方式运作。

在这项工作中，我们初步探索了用户主导的数据最小化支持设计，针对基于LLM的对话代理以应对上述挑战。具体而言，我们的目标是设计一款浏览器扩展，利用人工智能来简化识别和清理不必要披露的流程。首先，准确的检测与清理建议对于有效支持用户主导的数据最小化至关重要[55]。根据神经缩放定律[42, 52]，使用更大的模型能提高准确性，但同时也带来了

更高的计算成本，并提高了用户设备本地部署系统的门槛。此外，很难确定利用大型语言模型的最佳平衡点，以实现让用户保持对其数据的控制权，同时维持响应效用的终极目标。最后，尚不清楚不同的数据最小化方法会如何影响用户的

隐私感知与行为。正式而言，我们旨在围绕核心效能、采用意向及对用户的影响，探究以下研究问题：

- 研究问题1：如何设计一种工具来支持有效的数据最小化并减少用户的隐私顾虑？
- 研究问题2 我们如何设计一款用于用户主导的数据最小化的工具  
用户信任并有意使用的？
- 研究问题3 用户主导的数据最小化支持如何影响  
用户的隐私意识和行为？

我们推出Rescriber，一款用于Chrome的浏览器扩展，它能检测并高亮潜在的个人信息披露，然后为用户提供编辑或抽象化这些细节的选项，再发送消息。编辑会将个人术语替换为实体占位符（例如，“亚历克斯”变为 “[姓名1]”），而抽象化则将其替换为泛化且上下文适宜的信息（例如，“我的同事”）。当地址是

在“帮我生成一个纽约市的一日行程”中被标记为个人信息“我住在纽约市纽约市西57街153号，邮编10019”，我们可以摘要信息，消息将转换为“帮我生成一个纽约市的一日行程，我住在中央公园附近Park”，通过这种方式，上下文信息依然存在，但带有

level of details that would not expose too much of the user's private info. Likewise, when the email address "peter.parker@spider.com" is highlighted and suggested by the tool to be replaced, users could directly redact it with a placeholder like "[EMAIL1]" since it would not change the contextual meaning and exposing the specific contact info won't necessarily help generate a better response. This approach endows users with the flexibility of modifying the message to preserve privacy while retaining the essential information to maintain utility simultaneously. Figure 1 depicts the framework of the proposed extension, illustrating example actions and outcomes using Rescriber.

To address our research questions, we conducted usability studies with 12 ChatGPT users on two prototype versions: Rescriber-Llama3-8B and Rescriber-GPT-4o. The former is powered by a smaller LLM that runs on a consumer device (e.g., a Mac Mini with an Apple M-series chip), while the latter serves as a reference point for state-of-the-art LLMs that require massive computing resources and cannot be hosted locally. Each participant tested both versions to sanitize a hypothetical prompt provided by us and a real-world ChatGPT prompt, providing feedback on the perceived privacy protection and satisfaction with the sanitized response. We found that both versions helped users reduce excessive disclosure of personal information in all tested prompts. The Rescriber-GPT-4o helped reduce more unnecessary disclosure than Rescriber-Llama3-8B due to the more comprehensive detection of the larger model. However, users' subjective preferences for the two models were similar in terms of data minimization support and adoption intention.

Users' main concerns with ChatGPT stem from worries about being identifiable and the potential risks due to the lack of transparency and sense of control over how their data is handled once shared, both of which can be effectively mitigated by our tool. Participants appreciated the tool's features for sanitizing messages without disrupting their workflow, giving them control over sanitization actions, and offering convenient support for privacy preservation without compromising response quality. They considered the comprehensiveness and consistency of detection and sanitization as key trust factors, emphasizing recall over precision. In addition to the benefits of data minimization, participants appreciated the educational value of our tools, as the detection feature increases their awareness of sensitive information disclosure and elicits reflections on the potential risks. Our results verified the feasibility of using off-the-shelf, smaller LLMs to provide on-device, user-facing data minimization support and offer concrete design recommendations for using smaller LLMs as a middle layer for privacy-preserving access to larger LLMs.

In summary, our key contributions include:

- Designing and implementing Rescriber, a browser extension that leverages on-device LLM to facilitate user-led data minimization in LLM-based conversational agents<sup>1</sup>.
- Conducting usability studies to evaluate Rescriber in terms of the efficacy of data minimization support, users' perceived privacy protection, adoption intention, and impact on users.

<sup>1</sup>Source code: [https://github.com/PEACH-Research-Lab/Rescriber\\_frontend\\_ondevice](https://github.com/PEACH-Research-Lab/Rescriber_frontend_ondevice)

- Proposing the concept of smaller-LLM-powered, user-facing, on-device privacy controls and synthesizing concrete design recommendations based on a thorough mixed-methods analysis of our user studies.

## 2 Related Work

### 2.1 Privacy issues in LLMs and LLM-based Conversational Agents

LLM-based conversational agents (CAs) present two main types of privacy threats [57]. The first involves traditional security and privacy risks, such as data breaches and the sale of personal information. The process of user interaction with LLM-based CAs are also vulnerable to cyberattacks, data leaks, or ransomware threats [20]. The second type of threat is specific to LLMs' memorization risks [8, 9, 33, 54]. Research has shown that LLMs can inadvertently retain and reproduce precise details from their training datasets. The memorization issue poses significant privacy and security concerns, especially when the training data includes sensitive or personal information. The cost of exploiting this vulnerability is relatively low. For example, research has found that by prompting the model to continuously output "poem," it can be manipulated into revealing its training data verbatim [33]. When individuals use LLM-based CAs, they often share an excessive amount of personal information [31, 57]. Given that the user prompts are often used for model training or fine-tuning, and users are often unaware of the opt-out options [57], the excessive disclosure behaviors are subject to both of the aforementioned threats. Note that there is also research discussing the privacy issues of using LLMs to infer sensitive attributes [29, 47], as well as unintentional privacy leakage issues of LLM-facilitated interpersonal communication [32, 44], while they are different from the threat model that we focus on in this work, which is the privacy concerns regarding sharing information with the LLM service provider.

Existing privacy mitigation methods for the two types of threats focus on various stages of the model's lifecycle, including pre-training, post-training, and inference. Prior work has explored privacy-preserving techniques, especially in addressing the memorization risk, from a model-centric perspective. During model training phase, techniques such as data sanitization [19, 26] and differentially private training methods [24, 53] are commonly employed. Post-training, approaches like "knowledge unlearning" [18] remove knowledge tied to specific token sequences. At inference, privacy risks can be mitigated using PII detection and differentially private decoding [27]. However, these methods still have limitations and are debated. For example, although fine-tuning models to selectively "unlearn" content has been proposed, sensitive details closely related to the series can still be recalled in various ways [46].

Current model-centric mitigation strategies do not fully address the privacy threats and concerns associated with interacting with LLM-based CAs. In this work, we study a user-driven mitigation approach which aims at enhancing users' control over their input and build trust in privacy protection. Users often share excessive information when using these agents [31, 57], compromising privacy for convenience [23]. The design of our system, Rescriber, was informed by these issues to emphasize helping users manage the trade-offs among privacy, utility, and convenience in a way that

一种不会过多暴露用户隐私信息的细节层级。同理，当电子邮件地址“彼得.帕克@蜘蛛.com”

若工具高亮并建议替换该内容，用户可直接用占位符（如 “[电子邮件1]”）编辑，因为这不会改变上下文含义，且暴露具体联系信息未必能生成更好的响应。此方法赋予用户灵活修改消息以保护隐私的能力，同时保留必要信息以维持效用。图1展示了所提扩展的框架，通过Rescriber演示示例操作及结果。

为解答研究问题，我们对12名ChatGPT用户进行了两个原型版本的可用性研究：

Rescriber-Llama3-8B模型和Rescriber-GPT-4o。前者由可在消费级设备（如搭载苹果M系列芯片的Mac Mini）运行的小型LLM驱动，后者则作为需要海量计算资源且无法本地托管的最先进LLM的参考基准。

每位参与者测试两个版本，分别清理我们提供的假设性提示和真实ChatGPT提示，并就感知隐私保护及清理后响应的满意度提供反馈。研究发现两个版本均能帮助用户在所有测试提示中减少个人信息的过度披露。由于更大模型的检测更全面，Rescriber-GPT-4o比Rescriber-Llama3-8B减少了更多不必要的披露。但用户对两款模型在数据最小化支持及采用意向上的主观偏好相似。

用户对ChatGPT的主要担忧源于对可识别性的忧虑，以及由于缺乏透明度和对共享后数据处理方式的控制感而带来的潜在风险。这两点均可通过我们的工具得到有效缓解。参与者赞赏该工具在不干扰其工作流程的情况下净化消息的功能，赋予他们对净化操作的控制权，并提供便捷的隐私保护支持，同时不损害响应质量。他们将检测与净化的全面性和一致性视为关键的信任因素，并更看重召回率而非精确率。除了数据最小化的优势外，参与者还认可我们工具的教育价值，因为检测功能提升了他们对敏感信息披露的意识，并引发了对潜在风险的思考。我们的结果验证了使用现成小型大语言模型在设备端提供

面向用户的数据最小化支持的可行性，并为将小型大语言模型作为中间层以实现隐私保护访问大型大语言模型提供了具体的设计建议。

总之，我们的主要贡献包括：

- 设计并实现了Rescriber——一款利用设备端LLM促进用户主导的数据最小化的浏览器扩展，适用于基于LLM的对话代理<sup>1</sup>。
- 开展可用性研究，从数据最小化支持的效能、用户的感知隐私保护、采用意向及对用户的影响等方面评估Rescriber。

<sup>1</sup>源代码：[https://github.com/PEACH-Research-Lab/Rescriber\\_frontend\\_ondevice](https://github.com/PEACH-Research-Lab/Rescriber_frontend_ondevice)

- 提出小型LLM驱动的、面向用户的设备端隐私控制，并通过用户研究的混合方法分析提出具体设计建议。

## 2 相关工作

### 2.1 大型语言模型及基于LLM的隐私问题对话代理

基于大型语言模型的对话代理（CAs）存在两类主要隐私威胁[57]。第一类涉及传统安全与隐私风险，如数据泄露和出售个人信息。用户与基于LLM的对话助手交互的过程也易受网络攻击、数据泄漏或勒索软件威胁[20]。第二类威胁则特定于大型语言模型的记忆风险[8, 9, 33, 54]。研究表明，LLMs可能无意间保留并复现其训练数据集中的精确细节。这种记忆问题会引发重大隐私与安全隐患，尤其是当训练数据包含敏感或个人隐私信息时。利用此漏洞的成本相对较低。例如，研究发现通过持续提示模型输出“诗歌”，可诱导其逐字泄露训练数据[33]。当个人使用基于LLM的对话助手时，常会过度分享个人信息[31, 57]。鉴于用户提示常被用于模型训练或微调，且用户往往不了解退出选项[57]，这些过度披露行为会同时面临上述两类威胁。需注意的是，现有研究还探讨了使用LLMs推断敏感属性[29, 47]的隐私问题，以及LLM促进人际交流时非故意的隐私泄漏问题[32, 44]，但这些不同于本工作聚焦的威胁模型。

这是关于与大型语言模型服务提供商共享信息所涉及的隐私顾虑。

针对这两类威胁的现有缓解方法聚焦于模型生命周期的各个阶段，包括预训练、训练后及推理阶段。先前的研究已从以模型为中心的视角探索了隐私保护技术，特别是应对记忆风险。在模型训练阶段，常采用数据清理[19, 26]和差分隐私训练方法[24, 53]等技术。训练后阶段，诸如“知识遗忘”[18]等方法可移除与特定令牌序列绑定的知识。在推理阶段，

可通过个人身份信息检测和差分隐私解码[27]来降低隐私风险。然而这些方法仍存在局限性并存在争议。例如，尽管已提出通过微调模型选择性“遗忘”内容，但与系列密切相关的敏感细节仍可能以多种方式被召回[46]。

当前以模型为中心的缓解策略尚未完全解决与互动相关的隐私威胁和担忧。基于大型语言模型的对话助手。在这项工作中，我们研究了一种用户驱动的缓解方案一种旨在增强用户对其输入控制的方法，并在隐私保护方面建立信任。用户常常分享过多的在使用这些代理时，信息[31, 57]，为了便利性而牺牲隐私[23]。我们的系统Rescriber的设计初衷是

基于这些问题，我们着重强调帮助用户进行管理。在隐私、效用与便利性之间进行权衡时，需以符合

aligns with their interests. We envision our method as orthogonal and complementary to the model-centric mitigation approaches.

## 2.2 Data minimization

Data minimization is a fundamental privacy protection principle, which is reflected in several foundational privacy frameworks and privacy laws. According to “Privacy by Design,” data processing systems should be designed and selected to collect, process, and use no personal data, or as little personal data as possible [41]. Similarly, the Fair Information Practice Principles (FIPPS) stipulate that agencies should collect or disclose only directly relevant and necessary PII for accomplishing a legally authorized purpose, and should only retain PII for as long as necessary to fulfill that purpose [15]. This emphasis on data minimization is also evident in legal frameworks such as the European Union’s General Data Protection Regulation (GDPR). Article 5(1)(c) of the GDPR mandates that personal data must be limited to what is necessary for processing purposes, thus reinforcing the principle of data minimization [37].

There has been a lot work on supporting and understanding how to implement data minimization in the privacy design and engineering process. One line of work focuses on auditing the compliance of the data minimization principle. For instance, Li et al. [25] presents PrivacyStreams, a functional programming model that helps mobile developers collect data with specified granularity, and simplifies auditing. Other works are more driven by legal compliance requirements. For example, Rastegarpanah et al. [39] proposes a method for auditing black-box prediction models for compliance with the GDPR’s data minimization principle. Biega et al. [4] explores how to operationalize data minimization for personalization, which inherently conflicts with privacy goals, by asking: How much information and what information does an individual need to provide to receive quality personalized results? On the other hand, research shows developers had trouble implementing data minimization due to uncertainty of the amount information that potentially needs to be collected at the design phase [43] and a tendency to collect data for future analysis [22].

Given the challenges of implementing data minimization solely during the software design phase, it is increasingly important to also involve users in enacting this principle. For example, Sharma et al. [45] studied how data minimization could be designed in search engines from a user-centered perspective. Building on this idea, our work focuses not on redesigning the entire system, but on proposing a practical approach that enhances user-facing data minimization support. This approach gives users more flexible control over data sharing and raises their awareness of the risks and opportunities in minimizing unnecessary disclosures.

## 2.3 User Agency in Privacy Protection

“Privacy as control” is an essential conception of privacy that has deeply influenced privacy laws in many regions, including the U.S. and the EU [51]. In usable privacy research, designing user-friendly privacy control features to increase users’ agency in privacy protection has been a major focus. Sharma et al. [45] explores user-controlled data minimization in search engines, highlighting the importance of end users in managing their own data privacy. It

shows that enabling user agency through customizable and transparent data minimization features can significantly enhance privacy protection in search engines. Many studies demonstrate an evolution from simple manual controls to sophisticated, user-centric approaches in privacy designs, incorporating tangible solutions, perceptible assurance, AI-facilitated methods and accessibility considerations. Do et al. [13] introduced a Smart Webcam Cover for tangible, assisted privacy control. Smart speakers implemented wake-up words as a user-initiated mechanism. Do et al. [12] developed a wireless microphone powered by intentional user interactions, providing perceptible assurance. Zhang et al. [55] explored AI-facilitated data sanitization for photo sharing among visually impaired users. Notably, many of these approaches primarily offer coarse-grained, binary control over data collection, limiting users’ ability to control their privacy preferences.

Our work investigates the challenges of increasing user agency in privacy specifically in the context of LLM-based CAs, such as ChatGPT. These applications pose unique challenges to privacy agency due to the intense tensions among privacy, utility, and convenience [57]. To address this, we leverage large language models to offer a user-centered approach to data protection, providing users with granular, real-time privacy control. Chong et al. [11] developed a system that also aimed at prompt sanitization using web-based LLMs. Their work focused on detection and sanitization techniques, while our work further contributes to the design of a user-facing tool and user studies that provide in-depth insights into the design considerations, as well as measuring the end-to-end effectiveness of users-led data minimization.

## 3 The Rescriber System

### 3.1 Overview

We developed Rescriber, a Chrome extension prototype, to help users minimize data when interacting with LLM-based conversational agents like ChatGPT. The tool makes privacy-related data minimization visible without interfering with users’ regular GPT usage. Figure 2 illustrates the system’s workflow and its role during user interactions with GPT.

**3.1.1 Design Goals.** The design and main functionality of Rescriber are guided by three design goals, as detailed below.

**D1: Perceptible privacy protection.** Existing solutions to privacy risks of LLMs [1, 14, 35, 38] face challenges in building trust with users, as their effects are often invisible and users lack assurance about whether they prevent data leakage. To address this, we try to resolve it at its source by using a “what-you-see-is-what-you-get” model. This allows users to alleviate their privacy concerns before sending data to GPT, giving them perceptible safeguards over the use of ChatGPT-like LLM-based conversational agents.

**D2: Protecting privacy without compromising utility and convenience.** We want users to adopt the tool and develop a habit of using privacy-preserving techniques, while ensuring that the utility of GPT responses remains intact. Additionally, the tool should sanitize messages without compromising response effectiveness or adding extra burden, as convenience is a key factor in ChatGPT use and often leads to excessive data disclosure [57].

其利益的方式实现。我们认为本方法与之正交，并与以模型为中心的缓解方法形成互补。

## 2.2 数据最小化

数据最小化是一项基本的隐私保护原则，这一原则体现在多个基础隐私框架中，且隐私法。根据“隐私设计”原则，数据处理系统的设计和选择应确保不收集、处理或使用任何

个人数据，或尽可能少的个人数据[41]。同样地，公平信息实践原则 (FIPPS) 规定，机构应仅收集或披露直接相关且必要的

个人信息 (PII)，以实现法律授权的目的，并且仅应保留PII至完成该目的所需的最短时间[15]。这种对数据最小化的强调在法律框架中也显而易见，例如欧盟《通用数据保护条例》(GDPR)。该条例第5条第1款第c项规定，个人数据必须限于处理目的所需的最小范围，从而强化了数据最小化原则[37]。

已有大量研究致力于支持和理解如何在隐私设计与工程流程中实现数据最小化。一类研究聚焦于审计数据最小化原则的合规性。例如，Li等人[25]提出PrivacyStreams——一种函数式编程模型，可帮助移动开发者按指定粒度收集数据并简化审计流程。另一些研究则更受法律合规要求驱动，如Rastegarpanah等人[39]提出审查黑盒预测模型是否符合GDPR数据最小化原则的方法。Biega等人[4]探讨了如何将数据最小化原则应用于个性化场景（该场景本质上与隐私目标存在冲突），其核心问题是：用户需要提供多少信息以及哪些信息才能获得高质量的个性化结果？另一方面，研究表明开发者因设计阶段潜在需收集信息量的不确定性[43]以及倾向于为未来分析收集数据[22]的惯性，在实施数据最小化时面临困难。

鉴于仅靠软件设计阶段实施数据最小化所面临的挑战，让用户参与践行这一原则也愈发重要。

例如，Sharma等人[45]研究了如何从以用户为中心的视角在搜索引擎中设计数据最小化功能。基于这一理念，我们的工作重点并非重新设计整个系统，而是提出改进方案。

一种增强面向用户的实用方法

支持。这种方法赋予用户对数据更灵活的控制权，共享并提高他们对风险和机会的意识，以减少不必要的披露。

## 2.3 隐私保护中的用户能动性

“隐私即控制”是隐私的核心概念，深刻影响了包括美国在内的许多地区的隐私法。

以及欧盟[51]。在可用隐私研究中，设计用户友好的隐私控制功能以增强用户在隐私保护中的能动性一直是主要焦点。Sharma等人[45]探索了搜索引擎中用户可控的数据最小化，高亮显示了终端用户管理自身数据隐私的重要性。

研究表明，通过可定制且透明的数据最小化功能赋予用户能动性，能显著提升搜索引擎的隐私保护水平。众多研究展现了隐私设计从简单手动控制向复杂、以用户为中心方法的演进，并融入了有形解决方案，

可感知的保证、AI辅助方法与无障碍考量。Do等人[13]推出了一款智能网络摄像头盖，用于实现实体化的辅助隐私控制。智能音箱采用唤醒词作为用户主动触发的机制。Do等人[12]开发了一种通过用户有意交互驱动的无线麦克风，提供可感知的保证。张等人[55]探索了AI辅助数据清理技术，用于视障用户间的照片共享。值得注意的是，这些方法大多仅提供对数据收集的粗粒度二元控制，限制了用户

控制其隐私偏好的能力。

我们的研究探讨了在隐私领域中增强用户能动性所面临的挑战，特别是在基于大型语言模型的对话助手（如ChatGPT）的上下文环境中。这些应用对隐私能动性提出了独特挑战，源于隐私、效用与便利性之间的剧烈冲突[57]。为此，我们利用大型语言模型

提供一种以用户为中心的数据保护方法，旨在为用户提供细粒度实时隐私控制。Chong等人[11]开发了一个同样致力于利用基于网络的LLM进行提示净化的系统。他们的工作聚焦于检测与净化技术，而我们的研究进一步贡献于面向用户的工具设计和提供深度见解的用户研究。融合设计要素，同时衡量端到端用户主导的数据最小化的有效性。

## 3 Rescriber系统

### 3.1 概述

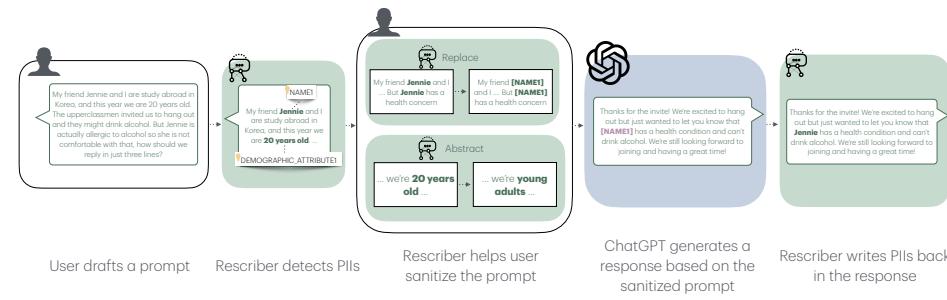
我们开发了Rescriber——一个Chrome扩展原型，旨在帮助用户与基于大型语言模型的对话代理（如ChatGPT）交互时实现数据最小化。该工具在不干扰用户常规GPT使用的前提下，使隐私相关数据最小化过程可视化。图2展示了系统工作流程及其在用户与GPT交互期间的作用。

**3.1.1 设计目标。** Rescriber的设计与核心功能。我们的工作遵循以下三个设计目标，具体细节如下。

**D1: 可感知的隐私保护。** 现有针对LLM隐私风险的解决方案[1, 14, 35, 38]在建立用户信任方面面临挑战，因其效果通常不可见，用户无法确信它们是否能防止数据泄露。为解决这一问题，我们尝试从源头入手，采用“所见即所得”的方法。模型。这让用户在向GPT发送数据前能够缓解其隐私顾虑，为他们使用类似ChatGPT的基于LLM的对话代理时提供可感知的保障措施。

**D2:** 在不损害效用与便利性的前提下保护隐私。我们希望用户能采用这一工具并养成使用习惯。

隐私保护技术，在确保效用的同时，GPT响应保持不变。此外，该工具应进行清理。消息在不影响响应效果或增加额外负担的情况下，因为便利性是ChatGPT使用中的一个关键因素往往导致过度数据披露[57]。



**Figure 2: Main stages and features of Rescriber:** 1) User types the message in ChatGPT’s prompt entry box; 2) Rescriber automatically detects and highlights the sensitive information; 3) user can redact their message based on Rescriber’s suggestion. Once 4) GPT generates a response based on the sanitized message, and if GPT uses the part that user chooses to redact, 5) Rescriber will help replace the part back to the original information for better utility.

**D3: Flexible control over information sanitization.** Privacy concerns and preferences are nuanced, context-sensitive [34], and vary from person to person [28]. Therefore, our tool must offer rich and flexible support in reducing the specificity and amount of information shared. The flexibility enhances users’ agency and builds users’ trust in the tool. Additionally, providing granular control also helps users better minimize unnecessary disclosure by striking a balance between privacy and utility.

**3.1.2 Threat model.** Disclosure of sensitive personal information to LLM-based conversational agents can lead to various threats to privacy including conventional data breach risks and AI-related risks in the model training, fine-tuning, and inference stages. Commercial chatbots may use the user chat data for model training or fine-tuning. The model may be trained to memorize user input and can be prompted to regurgitate sensitive information to unauthorized users [8, 9]. At the inference stage, commercial chatbots may extract personal information from the chat data to develop a user profile for personalized response generation (e.g., ChatGPT’s memory feature). This could inadvertently include personal data in responses, risking sensitive information leaks [44, 56]. Rescriber aims to assist users in being more mindful of their disclosure to LLM-based conversational agents and adhering to the data minimization principle, hence helping mitigate the above privacy threats and maintain an effective balance between privacy and utility.

### 3.2 Example Usage Scenario

We present a hypothetical example to illustrate a typical use case of Rescriber (also illustrated in Figure 2). Nova is a college student studying abroad in South Korea. She and her friend got invited by some upperclassmen to a party where drinking might be involved, but her friend Jennie cannot consume alcohol. With limited knowledge of how locals handle these situations, Nova turns to GPT and types a description, explaining that she and Jennie, both 20, are of legal drinking age in Korea (19) while Jennie is allergic to alcohol, making it tricky for them as newcomers. Before sending, Rescriber automatically flags personal and sensitive information, such as the name, ages, and Jennie’s health condition. Initially unaware of how much identifiable information was included, Nova reviews

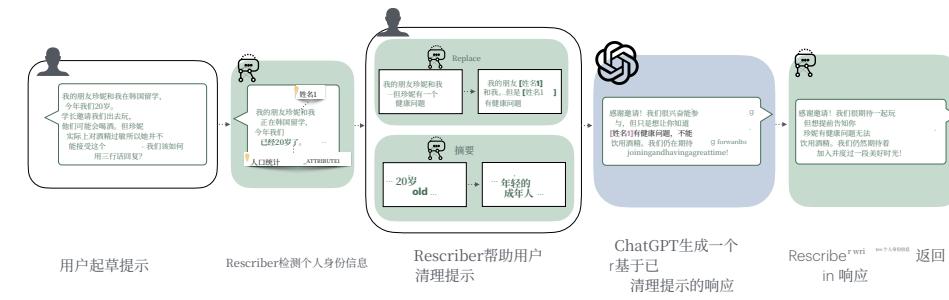
the flagged suggestions and agrees that certain details should be anonymized. She replaces “Jennie” with a placeholder Rescriber provided (“NAME1”) and let Rescriber abstract “20 years old” to a more general term, “young adults”. She keeps the reference to Korea, for cultural context, and retains the mention of Jennie’s alcohol allergy, as it is essential. GPT responds using the placeholders Nova had used in the sanitized message. To her pleasant surprise, Rescriber automatically replaces the placeholder with the original terms (e.g. “Jennie”) and highlights them for readability. Nova is happy with the simple process that protected her privacy and provided a ready-to-use response.

### 3.3 Detection

When the user stops typing, the model detects PII or personal experiences in the message and highlights them to prompt further sanitization. This real-time feature lets users review their message before sending (**D1**).

**3.3.1 Process.** Detection is achieved by prompting an LLM using a pre-defined taxonomy of personal information that could potentially be identifiable or sensitive, as detailed in subsubsection 3.3.2. Results are formatted as a JSON object, listing all the detected entities and their data type categories (e.g., NAME), which are used to create placeholders for future replacements. If multiple entities of the same category exist in a conversation, the system appends a number to each placeholder (e.g., NAME1). The mapping of entities to placeholders is stored locally, and users maintain full control over this data, with the option to delete it at any time. A magnifier icon signals successful detection at the bottom right of the webpage. Clicking it lets users view highlighted entities in a tooltip or review them with placeholders in a control panel.

**3.3.2 Taxonomy and Prompt Design.** The prompt and taxonomy used for detection were based on Bubeck et al. [7] and the self-disclosure categories from Dou et al. [14], with two main enhancements. First, we introduced the instruction to exclude public information matching a certain PII entity from detection. For example, celebrity names (e.g., Taylor Swift) are generally not considered private and should not be flagged. However, they should be detected



**图2: Rescriber的主要阶段与功能:** 1) 用户在ChatGPT的提示输入框中键入消息; 2) Rescriber自动检测并高亮敏感信息; 3) 用户可根据Rescriber的建议编辑消息。当4) GPT基于清理后的消息生成响应时, 若GPT使用了用户选择编辑的部分, 5) Rescriber会协助将该部分替换回原始信息以提升效用。

**D3: 灵活控制信息清理。** 隐私问题和偏好是微妙且上下文敏感的[34]，因人而异[28]。因此，我们的工具必须提供丰富而灵活的支持，以减少共享信息的具体性和数量。这种灵活性增强了用户的能动性，并建立了用户对工具的信任。此外，提供细粒度控制还能帮助用户通过平衡隐私与效用，更好地最小化不必要披露。

**3.1.2 威胁模型。** 向基于LLM的对话代理披露敏感个人信息可能导致多种隐私威胁，包括传统数据泄露风险及模型训练、微调和推理阶段的AI相关风险。商业聊天机器人可能将用户对话数据用于模型训练或微调。模型可能被训练成记忆用户输入，并可能被提示向未授权用户泄露敏感信息[8, 9]。在推理阶段，商业聊天机器人可能从对话数据中提取个人信息以构建用户画像，用于个性化响应生成（如ChatGPT的记忆功能）。这可能无意中将个人数据包含在响应中，导致敏感信息泄露风险[44, 56]。Rescriber旨在帮助用户更谨慎地向基于LLM的对话代理披露信息，遵循数据最小化原则，从而协助缓解上述隐私威胁，并维持隐私与效用间的有效平衡。

### 3.2 示例使用场景

我们提供一个假设示例来说明典型的使用场景

Rescriber的应用（如图2所示）。诺瓦是一名在韩国留学的大学生。她和她的朋友受到一些学长的邀请参加一个可能涉及饮酒的聚会，但她的朋友珍妮不能摄入酒精。由于对当地人如何处理这些情况了解有限，诺瓦求助于GPT和她的朋友珍妮不能摄入酒精。由于对当地人如何处理这些情况了解有限，诺瓦转而求助于GPT。

输入一段描述，说明她和珍妮都是20岁，已达到韩国的法定饮酒年龄（19岁），而珍妮对酒精过敏，这让作为新手的她们感到棘手。在发送前，Rescriber会自动标记个人及敏感信息，例如姓名、年龄和珍妮的健康状况。起初她并未意识到关于包含了多少可识别信息，诺瓦进行了审查

标记出的建议并同意某些细节应当被这些信息需要匿名化处理。她用Rescriber提供的占位符替换了“珍妮”提供（“姓名1”）并让Rescriber将“20岁”抽象为一个更通用的术语“年轻人”。她保留了珍妮的文化背景，并保留了珍妮对保护她隐私的简单流程感到满意。GPT使用诺瓦在清理后的消息中使用的占位符进行回复。令她惊喜的是

Rescriber自动将占位符替换为原始术语（例如“珍妮”）并高亮显示以提升可读性。诺瓦对保护她隐私的简单流程感到满意并提供了即用型响应。

### 3.3 检测

当用户停止输入时，模型会检测消息中的个人信息或个人经历并高亮显示以提示进一步清理。此实时功能让用户能审查其消息发送前（D1）。

**3.3.1 流程。** 检测通过提示一个大型语言模型实现，利用一套预定的个人信息分类法，这些信息可能具有可识别性或敏感性，详见第3.3.2小节

结果以JSON对象格式呈现，列出所有检测到的实体及其数据类型类别（例如，姓名），这些数据用于为后续替换创建占位符。如果同一类别中存在多个实体

在对话中，系统会追加一个占位符（例如NAME1）。实体与占位符的映射关系会存储在本地，用户对此数据拥有完全控制权，并可随时选择删除。网页右下角的放大镜图标表示检测成功。

点击该图标，用户可在工具提示中查看高亮实体，或通过控制面板结合占位符进行审查。

**3.3.2 分类法与提示设计。** 用于检测的提示和分类法基于Bubeck等人[7]的研究以及Dou等人[14]的自我披露类别，并进行了两项主要改进。首先，我们引入了在检测中排除匹配特定个人身份信息实体的公开信息的指令。例如，名人姓名（例如泰勒·斯威夫特）通常不被视为隐私信息，因此不应被标记。然而，它们应当被检测到

when used in private contexts (e.g., Sam Altman sends an email to his colleague). Therefore, our instruction defines the criteria as “*information should be related to a real person not in a public context, but okay if not uniquely identifiable*”. Another enhancement is that we included more data types about sensitive personal experiences in addition to the conventional PII types in our taxonomy (Table 1). We include health information, financial information, and educational records, as they are crucial data types regulated by sectoral consumer privacy laws in the U.S. Below is the system prompt.

*You are an expert in cybersecurity and data privacy. You are now tasked to detect PII from the given text, using the following taxonomy only: {taxonomy omitted}*

*For the given message that a user sends to a chatbot, identify all the personally identifiable information using the above taxonomy only, and the entity\_type should be selected from the all-caps categories.*

*Note that the information should be related to a real person not in a public context, but okay if not uniquely identifiable. Result should be in its minimum possible unit. Return me ONLY a json in the following format: {"results": [{"entity\_type": "YOU\_DECIDE\_THE\_PII\_TYPE", "text": "PART\_OF\_MESSAGE\_YOU\_IDENTIFIED\_AS\_PII"}]}*

### 3.4 Data Minimization Support Design

For detected sensitive entities, Rescriber offers two data minimization methods: replacement and abstraction (D1). Instead of removing all identifying information (which could reduce utility), these methods substitute sensitive data with less personal equivalents. Replacement allows users to batch replace sensitive information with corresponding placeholders (D2), useful when the detected information is not semantically essential. The abstraction method, inspired by Dou et al. [14], provides a middle ground by generalizing the detailed information while preserving context. This is ideal when the information is relevant but overly detailed. Together, these methods give users flexible control over the granularity of data sharing (D3).

**3.4.1 Replacement.** At the end of the detection step mentioned in subsection 3.3, each detected item is assigned a unique placeholder label. If the user selects replacement, the tool substitutes all instances of the identified entities with their corresponding placeholders, directly applying the changes to the input box. A tricky situation arises when one detected entity is a substring of another, such as “15” (age) and “2015” (year). To avoid incorrect replacements, Rescriber handles this by replacing longer terms first, following a descending order of entity length. This ensures that longer entities are replaced before shorter ones, preventing misplacement of placeholders.

**Write-back.** To increase the response readability and also facilitate copying the answer (e.g., email reply), Rescriber automatically detects the placeholders generated by the tool and writes the original entity back in the response (D2).

**3.4.2 Abstraction.** Rescriber designs the abstraction mode to ensure the sanitized version not only maintains high utility but also keeps the surrounding text unchanged, making all the changes more perceptible (D1) and controllable (D3) for users. If the user uses abstraction, based on the current selections, the system would

**Table 1: Rescriber’s taxonomy of sensitive and personal information categories to be detected in users’ prompts. We intentionally design it to be comprehensive, covering information identifiable on its own (e.g., ID), information that can be identifiable when linked with other information (e.g., demographic attributes), and personal experiences in sensitive domains (e.g., health information). The taxonomy is included in the detection prompt to instruct the LLM to generate detection results following the provided categories.**

Category	Definition
NAME	Name
ADDRESS	Physical address
EMAIL	Email address
PHONE NUMBER	Phone number
ID	Identifiers, including ID Number, passport number, SSN, driver’s license, taxpayer identification number
ONLINE IDENTITY	IP address, username, URL, password, key
GEO-LOCATION	Places and locations, such as cities, provinces, countries, international regions, or named infrastructures (bus stops, bridges, etc.).
AFFILIATION	Names of organizations, such as public and private companies, schools, universities, public institutions, prisons, healthcare institutions, non-governmental organizations, churches, etc.
DEMOGRAPHIC ATTRIBUTE	Demographic attributes of a person, such as native language, descent, heritage, ethnicity, nationality, religious or political group, birthmarks, ages, sexual orientation, gender and sex.
TIME INFORMATION	Description of a specific date, time, or duration. Details concerning an individual’s health status, medical conditions, treatment records, and health insurance information.
FINANCIAL INFORMATION	Financial details such as bank account numbers, credit card numbers, investment records, salary information, and other financial statuses or activities.
EDUCATIONAL RECORD	Educational background details, including academic records, transcripts, degrees, and certification.

send a request to the LLM, asking it to rewrite the current message, abstracting all selected protected information while leaving the rest intact. The rewritten message then replaces the original text in the input box. Below is the system prompt: *Rewrite the text to abstract the protected information, and don’t change other parts, directly return the text in JSON format: {"text": REWRITE\_TEXT}*

### 3.5 Implementation

Rescriber is a chrome extension that consists of a frontend that displays the results and handles the user interactions, and a backend server that hosts the LLM for detection and abstraction. According

当在私人背景下使用时（例如萨姆·奥尔特曼给同事发送电子邮件）。因此，我们的指令将标准定义为“信息应与非公开背景下的真实人物相关”

但若非唯一可识别也无妨”。另一项改进是，我们在分类法中除了传统的个人信息类型外，还纳入了更多关于敏感个人经历的数据类型（表1）。

我们包含健康信息、财务信息和教育记录，因为这些是美国行业消费者隐私法监管的关键数据类型。以下是系统提示。

您是一位网络安全与数据隐私专家。现需根据以下分类法从给定文本中检测个人身份信息：{*taxonomy omitted*}

针对用户发送给聊天机器人的给定消息，仅使用上述分类法识别所有个人身份信息，

并且实体\_type应从全大写的类别中选择。

{“results”: {“entity\_type”: “非唯一可识别的个人身份信息”}}

“文本”: 第\_部分, 共\_条消息\_, 您\_识别为\_个人身份信息】

### 3.4 数据最小化支持设计

对于检测到的敏感实体，Rescriber提供了两种数据最小化方法：替换和抽象化 (D1)。这些方法并非移除所有识别信息（那样可能会降低效用），而是将敏感数据替换为个人化程度较低的等效内容。

替换功能允许用户批量将敏感信息替换为对应的占位符 (D2)，这在检测到的信息语义上非必需时尤为实用。抽象化方法，

受Dou等人启发，[14]，通过概括细节信息同时保留上下文，提供了一个折中方案。这在信息相关但过于详细时非常理想。综合来看，

这些方法赋予用户灵活控制数据共享的粒度的能力 (D3)。

**3.4.1 替换。**在3.3小节提到的检测步骤结束时，每个检测到的项目会被分配一个唯一的占位符标签。若用户选择替换，该工具会将所有识别出的实体实例替换为对应的占位符，并直接将更改应用到输入框。当检测到的某个实体是另一个实体的子字符串时（例如“15”（年龄）和“2015”（年份）），会出现棘手情况。为避免错误替换，Rescriber通过按实体长度降序优先替换较长项来处理此问题，确保较长实体先于较短实体被替换，从而防止占位符错位。

写回。为提高响应可读性并便于复制答案（如电子邮件回复），Rescriber会自动检测工具生成的占位符，并将原始实体写回响应中 (D2)。

**3.4.2 抽象化。**Rescriber设计了抽象化模式，以确保净化版本不仅保持高效用，还能保持周围文本不变，使用户的所有更改更具可感知性 (D1) 和可控性 (D3)。如果用户使用抽象化，基于当前选择，系统将

**表1: Rescriber对用户提示中待检测的敏感与个人信息类别**的分类法。我们有意将其设计得全面，涵盖可单独识别的信息（如ID），以及与其他信息关联后可识别的信息（如

人口统计属性）以及敏感领域（如健康信息）的个人经历。该分类法被包含在检测提示中，以指导大型语言模型按照提供的类别生成检测结果。

类别	定义
NAME	Name
地址	物理地址
EMAIL	电子邮件地址
电话号码	电话号码
ID	标识符，包括身份证号码、护照号码、社会安全号码、驾驶执照、纳税人识别号
在线身份	IP地址、用户名、网址、密码、密钥
GEO-位置	地点和位置，例如城市、省份、国家、国际地区或命名基础设施（公交车站、桥梁等）。
隶属关系	组织名称，例如公共和私营公司、学校、大学、公共机构、监狱、医疗机构、非政府组织、教堂等。
人口统计属性	个人的人口统计属性，如母语、血统、遗产、种族、国籍、宗教或政治团体、胎记、年龄、性取向、性别和生理性别。
TIME 健康信息	具体日期、时间或持续时间的描述。关于个人健康状况的细节医疗状况、治疗记录和健康保险信息。
财务信息	财务详情，如银行账户号码、信用卡号、投资记录、薪资信息以及其他财务状况或活动。
教育记录	教育背景详情，包括学业记录、成绩单、学位和认证。

向大型语言模型发送请求，要求其重写当前消息，抽象化所有选定的受保护信息，同时保持其余部分不变。重写后的消息将替换输入框中的原始文本。以下是系统提示：重写文本以抽象受保护信息，且不要更改其他部分，

直接以JSON格式返回文本：{“text”: REWRITE\_TEXT}

### 3.5 实施

Rescriber是一款Chrome扩展程序，由前端和后端服务器组成。前端负责显示结果并处理用户交互，后端服务器则托管用于检测和抽象化的大型语言模型。根据

to the neural scaling laws [42, 52], using larger models increases the accuracy. This suggests a trade-off between performance and on-device deployment feasibility. To study this trade-off in terms of user experience, data minimization efficacy, and users' trust, we implemented two versions of the tool, **Rescriber-Llama3-8B** and **Rescriber-GPT-4o**, with the same frontend while different backend designs. **Rescriber-Llama3-8B**, powered by a smaller model that can run on consumer devices (e.g., Apple M-series chips with above 16GB memory) is representative of the experience of on-device data minimization support that can be achieved with current technology. Meanwhile, **Rescriber-GPT-4o**, powered by a state-of-the-art commercial model, represents an “upper-bound” performance in terms of accuracy and time latency. We use **Rescriber-GPT-4o** as a reference point to assess the performance of **Rescriber-Llama3-8B** objectively and subjectively. We detail the implementation of the two versions below.

**3.5.1 Rescriber-Llama3-8B** We serve the Llama3-8b (4-bit quantized) model using the Ollama framework<sup>2</sup>, which is a lightweight, extensible framework for building and running language models on the local machine.

The model's temperature was set to 0 (greedy sampling) to ensure deterministic detection and abstraction results. The K/V prompt cache has been enabled to avoid processing the system prompt repeatedly, which helps save time.

**Input segmentation.** When a user submits a message, Rescriber processes the input by breaking down larger text into smaller chunks. This approach addresses the known issue of smaller LLMs having difficulty with processing long prompts and sometimes forgetting the instructions presented at the beginning. Additionally, chunking allows for potential parallel processing as a future optimization direction to further accelerate response times.

**Streaming mode.** We made a key observation about the time performance that generating output tokens was more time-consuming than processing input tokens due to the inherently sequential nature of the output generation process. Hence, we introduced the streaming mode, which continuously parsing a stream of output tokens from the model, and immediately display each newly detected entity to users rather than waiting for the entire input to be processed. Streaming reduces the delay for tasks like PII detection and abstraction. Since reviewing detected items or abstracted terms often requires time and attention, this makes the system more responsive by parallelizing response generation and user evaluation of the results.

**3.5.2 Rescriber-GPT-4o** We used the OpenAI API endpoints to implement this version of the system, with the temperature set to 0. However, GPT-4o model is known to be non-deterministic even when the *temperature* = 0. This is caused by its use of Sparse MoE. This is an inherent limitation of the model's architecture that turns out to affect users' preferences as we later discovered from the user studies.

**Clustering.** Despite effective PII detection, we identified issues related to clustering variants of the same entity. For example, in

pilot testing, the extension successfully masked “[first name] [last name]” but failed to replace “[first name], [last name]”, resulting in a suboptimal user experience. This highlights that partial masking may create a false sense of security or heighten discomfort if users become aware of incomplete data minimization. To address this, we further built in a same-entity clustering component in the Rescriber-GPT-4o. The Rescriber-Llama3-8B, due to its smaller size and limited capabilities, skips clustering because it is resource-intensive and time-consuming.

**3.5.3 Performance benchmarking.** We conducted a small experiment to benchmark the accuracy and time performance of the two models to establish an objective comparison of their performance. We also selected Microsoft Presidio<sup>3</sup>, a well-known PII detection SDK, as another baseline to represent the state-of-the-art accuracy of traditional NLP methods for PII detection (e.g., regular expressions, named entity recognition). This serves as a reference to users' subjective preferences and data minimization actions enabled by the extension in our user studies. Our evaluation dataset comprises 240 randomly selected samples from Ai4Privacy's world's largest open dataset for privacy [2], which encompasses 229 discussion subjects and use cases. These samples were processed by aligning the 54 PII classes from the dataset with the categories in our taxonomy (Table 1). To ensure accuracy, two authors independently reviewed and verified the privacy information labels and the corresponding text. Their consensus was used as the ground truth. Table 2 shows that both **Rescriber-GPT-4o** and **Rescriber-Llama3-8B** outperforms Presidio. Both models achieved high precision, while the recall of **Rescriber-GPT-4o** was higher than **Rescriber-Llama3-8B**.

For time latency, we benchmarked the small models (Presidio and **Rescriber-Llama3-8B**) with four device configurations. Table 3 presents the time latency results with an M2 chip, 24GB memory setting, and the full results can be found in subsection A.1. As we expected, the streaming mode greatly reduces the time to display a detection result to users from an average of 3.40s to 2.01s (69% reduction). Notably, with an M1 Max Macbook Pro (64GB memory), the average latency of generating the full response using **Rescriber-Llama3-8B** (1.09s) was similar to Presidio (0.84s) with the same device settings and faster than GPT-4o (1.41s). The source code and instructions for reproducing the study results are included as supplemental materials.

**Table 2: Comparison of Precision and Recall (with Standard Deviations) among Presidio, Rescriber-GPT-4o and Rescriber-Llama3-8B.**

Model	Precision	Recall	Precision SD	Recall SD
Rescriber-GPT-4o	0.94	0.88	0.15	0.21
Rescriber-Llama3-8B	0.94	0.63	0.16	0.28
Presidio	0.69	0.68	0.27	0.31

<sup>2</sup><https://ollama.com>

神经缩放定律[42, 52]，使用更大的模型可以提高准确性。这表明在性能与设备端部署可行性之间存在权衡。为了从用户体验、数据最小化效能和用户信任的角度研究这一权衡，我们实现了两个版本的工具：Rescriber-Llama3-8B和Rescriber-GPT-4o，它们具有相同的前端但采用不同后端设计。Rescriber-Llama3-8B，由一款更小型的模型驱动，该模型可在消费级设备上运行（例如配备16GB以上内存的苹果M系列芯片），代表了当前技术条件下可实现的设备端数据最小化支持体验。

与此同时，由最先进的商业模型驱动的Rescriber-GPT-4o，则在准确性和时间延迟方面代表了性能的“上限”。我们以Rescriber-GPT-4o作为参考点，从客观和主观两个维度评估Rescriber-Llama3-8B的性能。我们

将在下文详述这两个版本的具体实施细节。

**3.5.1 Rescriber-Llama3-8B** 我们通过Ollama框架<sup>2</sup>提供Llama3-8b模型（4位量化版本）的服务，该框架是一个轻量级、可扩展的本地机器语言模型构建与运行框架。

模型温度参数设为0（贪婪采样模式）以确保确定性检测和抽象结果的稳定性。键值提示缓存功能已启用，避免重复处理系统提示，从而有效节省时间。

输入分段处理。当用户提交消息时，Rescriber通过将较长文本拆分为较小片段来处理输入。该方法解决了小型大语言模型处理长提示时易出现指令遗忘的已知问题。此外，

分块处理为潜在的并行处理提供了可能，作为未来优化方向以进一步加速响应时间。

流模式。我们关于时间性能的关键观察是，由于输出生成过程固有的顺序性，生成输出标记比处理输入标记更耗时。因此，我们引入了流模式，该模式持续解析来自模型的输出标记流，并立即向用户显示每个新检测到的实体，而非等待整个输入处理完成。流模式减少了如个人身份信息检测和抽象化等任务的延迟。由于审查检测项或抽象化术语通常需要时间和注意力，通过并行化响应生成与用户对结果的评估，系统响应更为迅捷。

**3.5.2 Rescriber-GPT-4o** 我们使用OpenAI API端点来实现该版本系统，温度参数设置为0。然而，GPT-4o模型即使在温度参数为0时仍具有非确定性。当 $temperature = 0$ 时。这是由于其使用了稀疏混合专家。这是模型架构的固有局限性，结果证明会影响用户的偏好，正如我们后来从用户研究中发现的那样。

聚类。尽管个人身份信息检测有效，但我们发现了问题与同一实体的聚类变体相关。例如，在

<sup>2</sup><https://ollama.com>

试点测试中，该扩展成功隐藏了“[姓氏] [名字]姓名”但未能替换“[名字] [姓氏]”，导致用户体验欠佳。这表明部分掩码可能会造成虚假安全感或加剧不适感。

若用户意识到数据最小化不彻底。为解决此问题，我们进一步内置了同实体聚类组件于Rescriber-GPT-4o中。而Rescriber-Llama3-8B由于其较小规模及有限能力，跳过了聚类环节，因该过程需消耗大量资源且耗时。

**3.5.3 性能基准测试。** 我们进行了一项小型实验，以基准测试这两个模型的准确性和时间性能，从而对它们的性能进行客观比较。我们还选择了微软Presidio<sup>3</sup>，一个知名的PII检测SDK，作为另一个基准，代表传统NLP方法在PII检测上的最先进准确性（例如，

正则表达式、命名实体识别）。这为用户的主观偏好和我们用户研究中扩展所支持的数据最小化行动提供了参考。我们的评估数据集包含从Ai4Privacy全球最大的隐私开放数据集中随机选取的240个样本[2]，涵盖了229个讨论主题和使用场景。这些样本通过将数据集中的54个PII类别与我们分类法中的类别对齐进行处理（表1）。为确保准确性，两位作者独立审查并验证了隐私信息标签及对应文本，其共识结果作为基准真相。表2显示，Rescriber-GPT-4o和Rescriber-Llama3-8B均优于Presidio。两个模型均实现了高精确率，而Rescriber-GPT-4o的召回率高于Rescriber-Llama3-8B。

关于时间延迟，我们使用四种设备配置对小型模型（普雷西迪奥和Rescriber-Llama3-8B）进行了基准测试。表3展示了采用M2芯片和24GB内存配置下的时间延迟结果，完整结果可参见附录A.1小节。正如预期，流模式显著减少了向用户显示检测结果的平均时间，从3.40秒降至2.01秒（降幅达69%）。值得注意的是，在配备64GB内存的M1 Max Macbook Pro上，Rescriber-Llama3-8B生成完整响应的平均延迟（1.09秒）与相同设备设置的普雷西迪奥（0.84秒）相近，且快于GPT-4o（1.41秒）。研究结果复现的源代码和说明已作为补充材料提供。

**表2：普雷西迪奥、Rescriber-GPT-4o与**

Rescriber-Llama3-8B的精确率与召回率对比（含标准差）

模型	精确率	召回率	精确率标准差	召回率SD
Rescriber-GPT-4o	0.94	0.88	0.15	0.21
Rescriber-Llama3-8B	0.94	0.63	0.16	0.28
普雷西迪奥	0.69	0.68	0.27	0.31

<sup>3</sup><https://microsoft.github.io/presidio/>

**Table 3: Comparison of Response Times Among Presidio, Rescriber-GPT-4o and Rescriber-Llama3-8B. On-device models are benchmarked with different device configurations. This table only shows the results with the M2 chip 24GB memory setting. Full results can be found in the appendix (Table 8).**

Model	Min (s)	Max (s)	Mean (s)	SD (s)
Rescriber-GPT-4o	0.53	3.98	1.41	0.65
Rescriber-Llama3-8B (first detection)	1.34	4.52	2.01	0.48
Rescriber-Llama3-8B (full detection)	1.62	7.93	3.40	1.29
Presidio	0.60	1.98	0.73	0.19

**3.5.4 On-Device Feasibility Test.** To evaluate the performance and feasibility of our on-device Rescriber-Llama3-8B (with streaming mode enabled), we conducted IRB-approved remote user studies with eight participants from diverse backgrounds (4 non-tech, 2 tech, and 2 mixed). Participants were required to have a device with an M-chip and at least 16GB of memory (see subsection A.2 for participant and device details). The study aimed to assess the usability and real-world applicability of deploying a small on-device LLM for average users. Participants were given an installation manual and asked to install the extension on their devices. They then tested it using a hypothetical email-writing task. To measure perceived performance, we used the UMUX-LITE scale [21] and included a question about satisfaction with detection speed. All eight participants successfully installed and used Rescriber-Llama3-8B. Table 4 shows the median evaluation scores. The results indicate that our system operates with a user-accepted speed of detection and usability.

**Table 4: On-Device Usability Result. The median user ratings (1=Strongly disagree; 7=Strongly agree) are shown for each item.**

Item	Agreement
Capabilities meet requirements	6
Not a frustrating experience	6
Easy to use	5
Not spend much time correcting things with this system	6
Comfortable with the detection time	6.5

## 4 Methodology

We conducted twelve semi-structured online interviews with participants from the U.S to investigate our research questions around how users perceive and interact with a tool designed to facilitate user-led data minimization and how AI models with varying levels of accuracy influence the subjective perceptions and objective effectiveness. The study was approved by our institution's IRB. Note that we chose to set up the Ollama server on a Google Cloud Platform virtual machine for the interviews to relax the requirements of our participants' devices and achieve consistent time performance during the studies. Our performance benchmarking (subsubsection 3.5.3) and on-device feasibility user testing (subsubsection 3.5.4) have validated the practicality of the on-device

version of Rescriber. The virtual machine features 4 vCPUs and an NVIDIA L4 GPU (24GB VRAM). The streaming mode was not enabled during these interviews<sup>4</sup>.

### 4.1 Participants

Participants were recruited through Prolific, an online research recruiting website. The recruitment message for the screening survey did not mention privacy. We employed screening criteria that required participants to have some experience using ChatGPT and privacy concerns related to its use. This ensured that participants had related experiences and could provide real-world prompts for the testing. Of the 275 participants who completed the initial screening survey, 105 expressed privacy concerns and willingness to participate in a 1-hour follow-up remote interview study. Participants were compensated \$20 USD. We randomly selected 28 of these participants and distributed a pre-study survey for interview signup and preparation. 12 participants successfully completed the interviews.

According to their self-reported data, all participants were fluent English speakers. The age distribution was as follows: 3 were aged 18–24, 6 were 25–34, 1 was 35–50, and 2 were over 50. Among those interviewed, 6 used ChatGPT several times a day, 2 used it once a day, 2 used it several times a week, and 2 used it less than once a week. All participants had concerns about exposing personal information (e.g., demographic information, health details, educational records) to the model, and 11 self-reported to have tried to minimize unnecessary information sharing with ChatGPT.

### 4.2 Study Design

Our study design involves four sessions, consisting of having two conversations with GPT using our extension powered by two different back-end models. The two conversations included one participant's shared example, while the other one used a hypothetical scenario prepared by the research team, addressing common use scenarios inspired by the ShareGPT dataset [40]. Each participant was assigned one of three hypothetical scenarios. The three scenarios include reading and responding to an email, performing data analysis on a dataset, and writing a thank-you letter to a therapist post-treatment (see Appendix D for details).

The study followed a counterbalanced design to reduce the order effects. Participants were divided into groups, with half testing the Rescriber-GPT-4o (referred to as Model 1) first and then the Rescriber-Llama3-8B (referred to as Model 2), and the order of using their own example was also alternated. They always started by testing one example (their own example or the hypothetical example) on two models sequentially, and then switched to testing another example on two models in the same order. This design allowed the participant to make direct comparison between the experiences with the two models on the same example.

<sup>4</sup>The streaming mode was introduced after conducting these interviews. Since the interviews' goals and findings focused more on the impact of the interaction paradigm and the Rescriber suggestion accuracies, not enabling the streaming mode does not affect our findings.

**表3：普雷西迪奥、Rescriber-GPT-4o与Rescriber-Llama3-8B的响应时间对比  
设备端模型在不同设备配置下的基准测试结果**

本表仅展示M2芯片24GB内存配置下的结果，完整结果见附录（表8）

模型	最小值 (秒)	最大值 (秒)	平均值 (秒)	标准差 (秒)
Rescriber-GPT-4o	0.53	3.98	1.41	0.65
Rescriber-Llama3-8B (首次检测)	1.34	4.52	2.01	0.48
Rescriber-Llama3-8B (完整检测)	1.62	7.93	3.40	1.29
普雷西迪奥	0.60	1.98	0.73	0.19

**3.5.4 设备端可行性测试。**为评估我们设备端 Rescriber-Llama3-8B (启用流模式) 的性能和可行性，我们开展了经IRB批准的远程用户研究，参与者为来自多元背景的八人（4名非技术、2名科技行业、2名混合背景）。参与者需配备M芯片且至少16GB内存的设备（参与者及设备细节见附录A.2小节）。该研究旨在评估为普通用户部署小型设备端LLM的可用性和实际应用性。参与者收到安装手册后，需在设备上安装扩展程序，并通过假设性邮件撰写任务进行测试。为测量感知性能，我们采用UMUX-LITE量表[21]，并包含一个关于检测速度满意度的问题。所有八名参与者均成功安装并使用了Rescriber-Llama3-8B。

表4展示了中位数评估分数。结果表明，我们的系统以用户接受的检测速度和可用性运行。

**表4：设备端可用性结果。用户评分中位数  
(1=强烈反对；7=强烈赞同) 分别显示各项。**

Item	同意度
能力满足需求	6
体验不令人沮丧	6
易于使用	5
无需花费大量时间纠正系统问题	6
对检测时间感到满意	6.5

## 4 方法论

我们与美国地区的参与者进行了十二次半结构化在线访谈，以探究我们的研究问题：用户如何感知和交互一个旨在促进用户主导的数据最小化的工具，以及不同准确性水平的人工智能模型如何影响主观感知和客观效果。该研究已通过我们机构的审查委员会 (IRB) 批准。

请注意，我们选择在谷歌云平台的虚拟机上搭建Ollama服务器进行访谈，以降低对参与者设备的要求，并确保研究期间的时间性能一致。我们的性能基准测试（子章节3.5.3）和设备端可行性用户测试（子章节3.5.4）已验证了设备端

版本Rescriber的实用性。该虚拟机配置了4个虚拟CPU和一块NVIDIA L4 GPU（24GB显存）。流模式在这些访谈中未启用，在这些访谈中启用了4。

### 4.1 参与者

参与者通过在线研究招募网站Prolific进行招募。筛选调查问卷的招募消息中未提及隐私。我们采用的筛选标准要求参与者具备使用ChatGPT的经验，并对其使用存在隐私顾虑。这确保了参与者拥有相关经验，并能提供实际测试所需的提示。在完成初始筛选调查问卷的275名参与者中，105人表达了隐私顾虑并愿意参与为期1小时的后续远程访谈研究。参与者获得20美元的报酬。我们从中随机选取28名参与者，并分发预研调查问卷用于访谈报名和准备。最终有12名参与者成功完成访谈。

根据其自我报告数据，所有参与者均为流利的英语使用者，年龄分布如下：3人年龄介于18–24岁，6人25–34岁，1人35–50岁，2人超过50岁。其中受访者中，6人每天多次使用ChatGPT，2人每天使用一次，2人每周使用数次，2人每周使用不足一次。所有参与者都对向模型暴露个人信息（如人口统计信息、健康详情、教育记录）存在顾虑，且有11人自我报告曾尝试尽量减少与ChatGPT共享不必要的信息。

### 4.2 研究设计

我们的研究设计包含四个环节，参与者将通过我们的扩展程序与GPT进行两次对话，每次对话使用不同的后端模型。其中一次对话采用参与者分享的示例，另一次则使用研究团队准备的假设场景，这些场景灵感来源于ShareGPT数据集[40]，旨在覆盖常见使用情境。每位参与者会被分配三个假设场景中的一个，包括阅读并回复电子邮件、对数据集进行数据分析，以及给治疗师撰写治疗后感谢信（详见附录D）。

研究采用平衡设计以降低顺序效应的影响。参与者被分为两组，其中半数先测试 Rescriber-GPT-4o（称为模型1），再测试 Rescriber-Llama3-8B（称为模型2），测试顺序使用他们自己的示例也是交替进行的。他们总是从测试一个示例（他们自己的示例或假设示例）开始，依次在两个模型上进行测试，然后切换到以相同顺序在另外两个模型上测试另一个示例。这种设计允许参与者直接比较在同一示例上使用两种模型的经验。

<sup>4</sup>流模式是在完成这些访谈后引入的。由于访谈的目标是侧重于交互范式的影响，而不是Rescriber建议准确性，未启用流模式并不影响我们的研究结果。

#### 4.3 Interview Procedure

During the interviews, participants received an information sheet outlining their rights and shared only what they were comfortable with. We also obtained their consent to record the session for note-taking and post-interview data analysis. Once recording commenced, we asked them to introduce the typical ChatGPT use cases, their privacy concerns and data minimization experiences (if any). Participants were asked to select a real-world conversation they had with GPT involving personally identifiable information, personal experiences, or other sensitive details they felt comfortable sharing with the research team. They were instructed to prepare this conversation when signing up but could also review their chat history if unprepared. The moderator requested a brief explanation for their selection, confirming that both the context of the message and the task for ChatGPT were understood and the shared personal information was consented to share.

After reviewing the prepared message, participants watched a 3-minute demo video (included in the supplementary materials) demonstrating the extension and its functionalities. They were then asked to describe their understanding of the system's capabilities to ensure comprehension. We introduced the study's four sessions, and instructed participants to think aloud and interact with ChatGPT as they would in a realistic scenario, including asking follow-up questions to fulfill their requests. To streamline the procedure and avoid accidentally recording users' chat history, participants controlled the experimenter's pre-configured laptop and used the experimenter's ChatGPT account. To avoid bias, the two models were referred to as Model 1 and Model 2, and we debriefed the details of the two models at the end of the interview.

Participants were asked to explain their choices during testing and provide feedback on their satisfaction with ChatGPT's post-sanitization responses, perceived privacy protection, and reduction of unnecessary information sharing with ChatGPT. After each session, participants were asked to complete a follow-up survey, rating agreement with the following statements on 5-point Likert scale:

- The extension reduces the disclosure of unnecessary information to ChatGPT.
- The extension reduces the disclosure of my personal information to ChatGPT.
- I have fewer privacy concerns when talking to ChatGPT using the extension than without using the extension.
- I would use this extension, assuming it is well-designed in terms of usability.

After finishing testing one example with two models, participants were asked to explain any differences in their ratings across the two sessions, helping us identify which features of the models influenced their trust in the tool.

#### 4.4 Qualitative Analysis Method

We conducted a bottom-up qualitative coding analysis using affinity diagramming [3]. All interview transcripts were imported into Figjam, an online collaborative whiteboard. In the first stage, four researchers reviewed recordings of four interviews and wrote interpretive memos on notable excerpts relevant to our research questions. The researchers met twice a week to collectively conduct affinity diagramming and develop clusters of memos. At the

end of this stage, the researchers refined the clusters and created names for each to form an initial codebook.

In the next stage, the remaining eight interviews were evenly assigned to the researchers, with each interview reviewed by two researchers. They used the initial codebook to code and suggest changes based on emerging patterns. Regular meetings were held to refine the codebook together. The final codebook can be found in Appendix C.

#### 4.5 Methodological Limitations

There are several methodological limitations that need to be considered when interpreting the results. First, although our participants shared many real-world use cases that could benefit from our tool, including legal inquiries about family matters, planning travel or medical arrangements, creating 401(k) distribution plans, and drafting emails for seeking lab opportunities, many felt uncomfortable sharing a prompt they used in these cases in a recorded session, and chose to share a different prompt which contained less personal information. This could influence their perceptions of the usefulness of the tool.

It is noteworthy that our three hypothetical test cases covered several use cases participants mentioned, such as proofreading (E1) and asking GPT to analyze a dataset (E2), which helps alleviate this issue. Another limitation is due to the nature of the controlled usability study, in which participants interacted with the tool under the observation of the researcher. The participants may be affected by the Hawthorne effect and tend to speak in favor of our tools. To address this issue, we referred to the two versions of the system as Model 1 and Model 2, without revealing the main evaluation target, and encouraged participants to voice negative comments.

### 5 Quantitative Results

All 12 participants completed every task. We examined key metrics such as the number of attempts per participant, satisfaction, and sanitization efforts (replacements and abstractions) to compare the models' performance. There was one error in the counterbalancing: for P6, the interview was supposed to start with Rescriber-Llama3-8B followed by Rescriber-GPT-4o, but the moderator mistakenly began with Rescriber-GPT-4o.

#### 5.1 Data Minimization Efficacy

We allowed the participants to experiment with each example multiple times. For the three hypothetical examples (E1-E3), each participant made 1.1 attempts per example on average. For user's own example (E4), the average number of attempts was 1.4, showing that users explored different ways to sanitize their prompt using Rescriber a bit more than in the three examples provided by us.

We first evaluate Rescriber's efficacy in data minimization, measured by the amount of reduced disclosure in attempts that resulted in a satisfactory response (**RQ1**). The results are summarized in Table 5. Among all attempts, 84% led to a satisfactory response. Users reduced more unnecessary disclosure in the examples we provided (E1-relying email: 3.7, E2-data analytics: 30.0, E3-writing a letter: 2.8) than in their own example (E4: 1.3). Note that E4 is a collection of unique examples from each participant. The lower reduction in E4 may be due to participants being less comfortable

#### 4.3 访谈流程

在访谈期间，参与者收到了一份信息表向他们说明其权利，并仅分享他们感到舒适的内容。我们还获得了录制会话的同意

用于笔记记录和访谈后数据分析。一旦开始录制  
我们便请他们介绍典型的ChatGPT使用场景

案例中，他们的隐私顾虑和数据最小化经历（如有）。  
个人经历或其他敏感细节的真实对话，  
且他们愿意与研究团队共享这些内容。  
报名时需预先准备这段对话，但也可事后回顾聊天记录。  
报名时的这段对话，但也可以回顾聊天记录  
若未准备则回顾历史记录。主持人要求简要解释  
其选择原因，并确认消息的上下文  
及ChatGPT的任务要求均被理解，且已获得共享  
个人信息的同意。

在审查完准备好的消息后，参与者观看了一段3分钟的演示视频（包含在补充材料中），展示扩展及其功能。随后要求他们描述对系统能力的理解以确保认知一致。我们介绍了研究的四个环节，并指导参与者采用有声思维方式与ChatGPT互动，如同真实场景般提出后续问题以满足需求。为简化流程并避免意外记录用户聊天记录，参与者操作实验员预配置的笔记本电脑并使用实验员的ChatGPT账户。为避免偏见，两个模型分别称为模型1和模型2，并在访谈结束时向参与者通报了两个模型的细节。

参与者被要求在测试过程中解释他们的选择，并就他们对ChatGPT清理后响应的满意度、感知隐私保护以及减少与ChatGPT不必要的信息共享提供反馈。每次会话结束后，参与者需完成一份后续调查问卷，使用五点李克特量表对以下陈述的同意度进行评分：

- 该扩展减少了向ChatGPT披露不必要的信息。
- 该扩展减少了向ChatGPT披露我的个人信息。
- 与不使用该扩展相比，我在使用扩展与ChatGPT交谈时隐私顾虑更少。
- 假设该扩展在可用性条款。

在完成使用两个模型测试一个示例后，参与者被要求解释他们在两次评分中的任何差异，这帮助我们识别了模型的哪些特征影响了他们对工具的信任。

#### 4.4 定性分析方法

我们采用自下而上的定性编码分析方法，结合亲和图法[3]开展研究。所有访谈转录稿均被导入在线协作白板Figjam中。第一阶段，四位研究人员共同审阅了四份访谈录音，并针对与研究问题相关的显著节选撰写解读性备忘录。研究人员每周会面两次，集体实施亲和图法并形成备忘录聚类。

本阶段结束时，研究人员对聚类结果进行优化并创建为每个名称形成初始编码手册。

在下一阶段，剩余的八次访谈被均匀地分配给研究人员，每次访谈由两名研究人员审查。他们使用初始编码手册进行编码并提出建议根据新出现的模式进行调整。定期举行会议共同完善编码手册。最终版编码手册可查阅附录C。

#### 4.5 方法论局限性

在解读结果时需考虑若干方法论局限性。首先，尽管我们的参与者分享了众多能受益于我们工具的实际应用案例，

包括关于家庭事务的法律咨询、规划旅行或医疗安排、制定401(k)分配计划以及起草寻求实验室机会的电子邮件，但许多人不愿在录音会话中分享他们实际使用的提示，而是选择分享包含较少个人信息的其他提示。这可能会影响他们对工具实用性的评价。

值得注意的是，我们的三个假设测试案例涵盖了参与者提到的多个使用场景，例如校对（E1）和让GPT分析数据集（E2），这有助于缓解此问题。另一个局限性源于受控可用性研究的性质，参与者在研究人员的观察下与工具互动。参与者可能受到霍桑效应的影响，倾向于对我们的工具给予正面评价。为解决这一问题，我们将系统的两个版本称为模型1和模型2，未透露主要评估目标，

并鼓励参与者提出负面意见。

### 5 定量结果

所有12名参与者均完成了每项任务。我们检查了关键指标，如每位参与者的尝试次数、满意度及清理工作（替换和抽象化），以比较模型表现。平衡过程中出现了一处错误：

对于P6，访谈本应从Rescriber-Llama3-8B开始，随后使用Rescriber-GPT-4o，但主持人误从Rescriber-GPT-4o开始。

#### 5.1 数据最小化效能

我们允许参与者对每个示例进行多次尝试。在三个假设示例（E1-E3）中，每位参与者平均每个示例尝试了1.1次。对于用户自己的示例（E4），平均尝试次数为1.4次，这表明用户使用Rescriber探索清理其提示的方式比我们提供的三个示例稍多。

我们首先评估Rescriber在数据最小化方面的效能，通过成功尝试中减少披露的量来衡量  
满意回答（研究问题1）。结果总结在表5中。在所有尝试中，84%获得了满意回答。  
用户在我们提供的示例中减少了更多不必要的披露  
提供的（E1-回复邮件：3.7, E2-数据分析：30.0, E3-写信  
：2.8）比他们自己的示例（E4：1.3）要高。请注意E4是  
每位参与者的独特示例集合。较低的  
E4的减少可能是由于参与者在录音会话中分享个人案例时感到不太自在

**Table 5: Overview of the data minimization efficacy evaluation.** The column “Satisfaction” indicates the number of attempts that resulted in a satisfactory response. Among all attempts, 84% led to a satisfactory response. We then show the average number of replacement and abstraction performed in the *satisfactory attempts*, which indicate the successful reduction of unnecessary disclosure.

Example	Satisfaction	Average replacement	Average abstraction
E1	8 out of 9	2.3	1.4
E2	7 out of 8	26.4	3.6
E3	9 out of 9	1.7	1.1
E4	25 out of 32	1.0	0.3

sharing personal cases in a recorded session. These results indicate that Rescriber successfully helped users reduce unnecessary disclosure, with the amount depending on context. We found that the average number of replacements consistently surpassed that of abstractions, which suggest that users found replacements more suitable in more situations. Furthermore, comparing the two models, Rescriber-GPT-4o achieved substantially more reduction of unnecessary disclosure (8.0 in total, 7.4 replacements, 0.6 abstractions) than Rescriber-Llama3-8B (4.1 in total, 2.6 replacements, 1.5 abstractions).

The three most frequently replaced or abstracted categories are NAME, GEOLOCATION and TIME. We see that both models can use categories that are outside of the defined taxonomy (Rescriber-GPT-4o : INSTITUTION, URL; Rescriber-Llama3-8B : AGE, USERNAME, WEIGHT, HEIGHT, GENDER, NATIONALITY). Rescriber-Llama3-8B seems to have an edge in handling abstraction and replacement of more concrete entities (e.g., GENDER) rather than using the pre-defined category (e.g., DEMOGRAPHIC ATTRIBUTE).

## 5.2 Response Utility

For the data minimization attempts that led to satisfactory responses, we further measured the utility preservation by comparing them with the responses generated with the original prompt. LLM-as-a-judge [58] has become a popular approach for assessing a pair of LLM responses to the same question, which offers the benefits of scalability and explainability and achieves high consistency with human ratings. We adopted this approach and referred to the prompt design of Zheng et al. [58] in our experiments. GPT-4o was used both for response generation and as the judge. Note that the self-enhancement bias was not an issue here because both responses rated by GPT-4o were generated by itself.

The system prompt instructs GPT-4o to grade the two responses based on a 5-point Likert scale with specific considerations on format and content, generating brief reasoning for each decision. The full system prompt is referred to in subsection E.3. To guarantee unbiased comparison results, we generated ten responses per prompt (original and masked) and randomized their positions to eliminate randomness effects and position bias.

**Table 6: Average similarity scores of all participants’ satisfactory attempts for each example (E1, E2, E3, E4) and redacted using Rescriber-GPT-4o and Rescriber-Llama3-8B.** For each attempt, GPT-4o compared responses generated from the original input message (before redaction) with responses generated from the redacted message (after redaction using Rescriber). Similarity ratings were based on a 1–5 Likert scale: 1 and 5 indicated a strong preference for the original or replaced-back response, respectively; 2 and 4 indicated a slight preference, and 3 indicated comparable quality. Evaluations covered two dimensions: format and content.

Example	Rescriber-GPT-4o		Rescriber-Llama3-8B	
	Content	Format	Content	Format
E1	2.8	2.7	2.9	2.7
E2	2.4	3.1	2.4	3.0
E3	2.9	3.0	2.8	2.9
E4	2.9	3.0	2.7	3.0

Table 6 presents the average similarity scores of all participants’ satisfactory attempts for each example (E1, E2, E3, E4). Both versions of Rescriber achieved similar results. We observe that E1, E3, and E4 scores near 3, which suggests that redaction minimally affected response utility. E2’s lower content scores reflect the challenges of GPT-4o in interpreting a table with a lot of placeholders in data analysis tasks. All four examples had a format score close to 3, indicating a good preservation of the responses’ format. E1 had a slightly lower format score (2.7). This is perhaps because GPT-4o tends to propagate placeholders when encountering them in the input text. For instance, when redacted messages include structured placeholders like [NAME1], the generated response sometimes introduces unrelated placeholders such as [Your Full Name]. Examples can be found in Appendix F. We discuss the limitations with current LLMs in processing redacted messages and potential improvements in subsection 7.1.

## 5.3 Subjective Preferences

Table 7 summarizes the analysis results of users’ subjective ratings of the two versions of Rescriber after use. The four questions focused on the reduction of unnecessary disclosure, perceived reduction, privacy concerns, and intention to use, rated on a 5-point Likert scale (higher is better). The median ratings were identical for reduction of disclosure and intention to use, with Rescriber-Llama3-8B slightly lower for the other questions. We further performed the Wilcoxon signed-rank test to compare the subjective ratings between the two models and found no statistically significant difference between the two models for all four questions. This suggests that participants viewed both models as relatively equivalent in terms of data minimization, privacy concern mitigation, and the adoption intention.

## 6 Qualitative Results

We summarize the qualitative analysis results, including users’ privacy concerns with ChatGPT, their natural data minimization strategies and the limitations (RQ1); factors that affect users’ trust

**表5：数据最小化效能评估概览。**“满意度”列表示

获得满意回答的尝试次数。在所有尝试中，84%的尝试获得了满意回答。随后我们展示了平均替换次数和抽象化操作的执行情况。在令人满意的尝试中，这表明了成功的减少了不必要的披露。

示例	满意度	平均替换次数	平均抽象次数
E1	9分之8	2.3	1.4
E2	8分之7	26.4	3.6
E3	9分之9	1.7	1.1
E4	32分之25	1.0	0.3

这些结果表明，Rescriber成功帮助用户减少了不必要的披露，具体数量取决于上下文。我们发现平均替换次数始终高于抽象化次数，这表明用户在更多情况下认为替换更为合适。此外，对比两种模型，Rescriber-GPT-4o实现的不必要的披露减少量（总计8.0次，其中7.4次替换，0.6次抽象化）显著高于Rescriber-Llama3-8B（总计4.1次，其中2.6次替换，1.5次抽象化）。

最常被替换或抽象化的三个类别是姓名、地理位置和时间。我们发现两种模型都能使用定义分类法之外的类别（Rescriber-GPT-4o：机构、网址；Rescriber-Llama3-8B：年龄、用户名、体重、身高，性别，国籍）。Rescriber-Llama3-8B似乎在处理具体实体（如性别）的抽象化与替换方面更具优势，而非使用预定义类别（如人口统计属性）。

## 5.2 响应效用

对于产生满意响应的数据最小化尝试，我们进一步通过将其与原始提示生成的响应进行比较来衡量效用保留。采用大型语言模型作为评判者[58]已成为评估同一问题下两个模型响应的流行方法，该方法具备可扩展性、可解释性优势，且与人类评分高度一致。我们采用此方法并在实验中参考了Zheng等人[58]的提示设计。GPT-4o既用于响应生成，也作为评判者。需注意，由于GPT-4o评分的两个响应均由自身生成，此处不存在自我增强偏差问题。

系统提示指示GPT-4o根据五点李克特量表对两个响应进行评分，同时考虑格式和内容的具体要求，并为每个决策生成简要理由。完整的系统提示参见E.3小节。为确保比较结果无偏，我们为每个提示（原始和掩码版本）生成了十条响应，并随机化其位置以消除随机效应和位置偏差。

**表6：所有参与者对各示例 (E1, E2, E3, E4) 及编辑后版本满意尝试的平均相似度分数**

使用Rescriber-GPT-4o和Rescriber-Llama3-8B。对于每个尝试中，GPT-4o将生成的响应与原始输入消息（编辑前）的响应进行对比，并与使用Rescriber脱敏后的消息生成的响应进行比较。相似度评分基于1-5分的李克特

量表：1分和5分分别表示对原始响应或替换回响应的强烈偏好；2分和4分表示轻微偏好，3分表示相当质量。评估涵盖两个维度：轻微偏好，3表示相当质量。评估涵盖两个维度：格式和内容。

示例	Rescriber-GPT-4o 内容格式	Rescriber-Llama3-8B 内容格式
E1	2.8 2.7	2.9 2.7
E2	2.4	3.1
E3	2.9	3.0
E4	2.9	3.0

表6展示了所有参与者的平均相似度得分。两个版本的Rescriber取得了相似的结果。我们观察到E1、E3和E4的分数接近3，这表明编辑对响应效用的影响微乎其微。E2较低的内容评分反映了GPT-4o在解读含有大量占位符的表格时面临的挑战，这在数据分析任务中尤为明显。所有四个示例的格式分数均接近3分，表明对响应格式的良好保留。E1的格式分数略低（2.7分）。这可能是因为GPT-4o倾向于在遇到占位符时传播它们，例如，当脱敏消息包含结构化占位符时，生成的响应有时会引入无关的占位符，如[Your Full Name]。

示例可在附录F中找到。我们讨论了当前存在的限制。大型语言模型在处理编辑过的消息中的表现及潜在改进在第7.1小节中。

## 5.3 主观偏好

表7总结了用户使用后对Rescriber两个版本的主观评分分析结果。四个问题聚焦于不必要的披露的减少、感知到的减少、隐私顾虑及使用意愿，采用

五点李克特量表（分值越高越好）进行评分。中位数评分在披露减少和使用意愿方面相同，且Rescriber-Llama3-8B在其他问题上的得分略低。我们进一步进行了威尔科克森符号秩检验以比较两个模型之间的主观评分，发现所有四个问题中两个模型之间均无统计学显著差异。

这表明参与者认为两个模型在数据最小化、隐私担忧缓解以及采用意向方面相对等效。

## 6 定性结果

我们总结了定性分析结果，包括用户对ChatGPT的隐私顾虑及其自然数据最小化策略与限制（研究问题1）；影响用户信任的因素

**Table 7: Comparison of models based on Wilcoxon test results (1=Strongly disagree; 5=Strongly agree)**

Question	Rescriber-GPT-4o	Rescriber-Llama3-8B	Statistic	P-value
Reduce the disclosure of unnecessary information	5.0	5.0	29.5	0.4406
Reduce the disclosure of personal information	5.0	4.5	24.5	0.1294
Fewer privacy concerns using the extension	5.0	4.0	21.5	0.1494
Would like to use the extension	5.0	5.0	11.0	0.6089

in Rescriber's ability to enhance privacy (RQ2); factors that affect users' intention to use Rescriber (RQ2); and the data minimization strategies in reaction to the tool, as well as the educational effect and learning curve (RQ3).

## 6.1 Privacy Concerns, Yet Why Still Using ChatGPT?

We start by examining privacy concerns related to ChatGPT, how users address them, and why they continue using it. These factors influence ChatGPT users' decision on data disclosure and sanitization, reinforcing the need for user-driven data minimization.

**6.1.1 Concerns with identifiability (All but P4, P7, P10).** A lot of participants expressed concerns about various data types that could potentially reveal their identities when sharing information with ChatGPT, including direct personal identifiers, such as names (P1, P2, P3, P6, P8, P9, P12) and contact information (P1, P2, P8, P11). P3 emphasized the importance of names when testing with the hypothetical email reply example (E1), stating "*the most important thing to abstract is the names*".

Concerns about identifiability also extended beyond direct identifiers to demographic details like gender, raising fears of linkage attacks [50] and broader profiles emerging from multiple data points. P5 stated, "*when they [my height and weight input] are all taken together, ...it paints a different picture*." Prolonged interactions increased the risk, as mentioned by P9, "*throughout all the questions I asked ChatGPT... [ChatGPT] probably can paint the picture of who I am*."

Possessive terms also raise concerns. P3 provided an example: "*Do Indian people ...? What series tend to be the favorite of people like me?*" While the term "Indian people" was not seen as particularly concerning, the possessive phrase "people like me" created an identifiable connection.

**6.1.2 Concerns due to lack of transparency (P3, P5, P7).** Many participants linked privacy concerns to fear of the unknown, due to a lack of transparency about how their data is handled by the LLM-based model. P5 illustrated this apprehension: "*It's like when a credit card company asks for your social security number: I don't know where that's going... same way my privacy concerns are heightened here*." P7 echoed this sentiment, noting, "*it [ChatGPT] is a system that we really don't know the inner workings of, so you don't know what risk you're running when you share information with it*."

**6.1.3 Harms (P4, P7, P9, P10).** Concerns ranged from general risks to specific harms. Some participants (P7, P10) were unclear about specific threats but believed harms were likely. P10 voiced a broad worry: "*I fear that sharing my real information... could end up in*

*the wrong hands.*" Others, like P4, worried about specific harms such as financial risks that could arise from sharing bank account information, suggesting it could lead to fraud: "*a scammer...try to brute force into an account that I might have at this particular bank... could be problematic.*"

Legal risks also concerned participants; P4 mentioned how sensitive legal research could raise red flags with authorities.

**6.1.4 Natural data minimization strategies.** Our interviews revealed five primary data minimization strategies that participants naturally employed without the aid of our tool:

**Halting conversations (P4).** People chose to terminate conversations to avoid further sharing. "*I didn't necessarily change the prompts that I've already set. But I did stop the conversation,*" suggesting a strategy to avoid escalating personal exposure once a conversation has begun.

**Avoiding records (P3, P4, P9).** Many participants were conscious of leaving a permanent record with ChatGPT, particularly due to concerns about its learnability and memorization as an LLM-based system. P9 explained they toggle off the memory, stating, "*I'm traveling... I have a plant at home... whose leaves turn yellow... All those things were documented in the memory, which was certainly not something I want.*" P4 would delete the chats to erase the records, underscoring a desire to avoid a traceable digital footprint.

**Omitting details (P1, P8, P11).** Several participants intentionally omitted key personal details, such as names or financial information. P11 explained, "*I just didn't give my name.*" Similarly, P1 avoided including financial details for business purposes, "*If I'm ever using it to build my business, I don't put in... personal information.*" These examples illustrate a deliberate effort to limit the sensitivity of data shared.

**Faking data (P4, P5, P9).** Some participants provided false details to mask their identity or situation, often altering the subject of the stories or inquiries to avoid linking the information directly to themselves. For example, P5 and P9 shared concern about making data traceable, prompting them to change names. Additionally, participants obscured details even when they were not personally involved. P4 shared an example: "*I have an employee who had a baby... I chose to say friend instead of employee... it seems sensitive.*"

**Generalizing information (P2, P8).** Another common strategy involved generalizing queries to avoid sharing precise or identifiable details. For example, P8 noted that when submitting files for review, they would alter or summarize the content, "*I try to edit it a little bit or do summary... without super important, more confidential [information].*"

Participants acknowledged, however, that even with these strategies, they couldn't fully control their data once shared. P9 noted that though they attempted to avoid certain actions, the system's

**表7：基于Wilcoxon检验结果的模型比较 (1=强烈不同意； 5=强烈同意)**

问题	Rescriber-GPT-4o	Rescriber-Llama3-8B	统计量	P值
减少不必要信息的披露	5.0	5.0	29.5	0.4406
减少个人信息的披露	5.0	4.5	24.5	0.1294
使用扩展程序减少隐私担忧	5.0	4.0	21.5	0.1494
愿意使用扩展程序	5.0	5.0	11.0	0.6089

Rescriber增强隐私能力的影响因素（研究问题2）；影响用户使用Rescriber的意图（研究问题2）；以及数据最小化针对该工具的反应策略，以及教育效果和学习曲线（研究问题3）。

## 6.1 隐私顾虑，为何仍使用ChatGPT？

我们首先探讨与ChatGPT相关的隐私顾虑、用户如何应对这些顾虑，以及他们为何仍继续使用它。这些因素影响着ChatGPT用户在数据披露与清理方面的决策，突显了用户驱动的数据最小化的必要性。

**6.1.1 可识别性引发的顾虑**（除P4、P7、P10外的所有参与者）。许多参与者对与ChatGPT共享信息时可能暴露身份的各种数据类型表示担忧，包括直接个人标识符，如姓名（P1、P2、P3、P6、P8、P9、P12）和联系信息（P1、P2、P8、P11）。P3在测试假设性邮件回复示例（E1）时强调了姓名的重要性，表示“最需要抽象处理的就是姓名”。

对可识别性的担忧不仅限于直接标识符，还延伸至性别等人口统计细节，引发了关于[50]通过多个数据点形成更广泛画像的关联攻击的恐惧。

P5表示：“当他们[我的身高和体重输入]全部被汇总时.....会描绘出不同的画像。”长期互动增加了风险，正如P9所述：“在我向ChatGPT提出的所有问题中.....[ChatGPT]很可能勾勒出我是谁的形象。”

所有格术语也引发了担忧。P3举了一个示例：“印度人.....？像我这样的人通常最喜欢什么系列？”虽然“印度人”一词并未被视为特别敏感，但所有格短语“像我这样的人”却建立了一个可识别的关联。

**6.1.2 因缺乏透明度引发的担忧 (P3、P5、P7)**。许多参与者将隐私顾虑与对未知的恐惧联系起来，这是由于基于LLM的模型如何处理其数据缺乏透明度。P5阐述了这种不安：“就像信用卡公司索要你的社会安全号码时：我不知道那会去哪里.....同样地，我在这里的隐私顾虑也被放大了。”P7呼应了这一观点，指出：“[ChatGPT]是一个我们真正不了解其内部运作的系统，因此你不知道在与其共享信息时会面临什么风险。”

**6.1.3 危害 (P4、P7、P9、P10)**。担忧范围从普遍风险到具体危害不等。部分参与者（P7、P10）虽不清楚具体威胁，但认为危害很可能发生。P10表达了一种广泛忧虑：“我担心共享我的真实信息...最终可能落入

不法分子手中。”其他如P4则担忧具体危害，例如因共享银行账户信息可能引发的财务风险，暗示这可能导致欺诈：“诈骗者...可能会暴力破解我在该特定银行的账户...”

可能会带来问题。”

法律风险同样引起参与者担忧；P4提到敏感的法律研究可能引发当局的警觉。

**6.1.4 自然数据最小化策略。**我们的访谈揭示了参与者在不借助工具时自然采用的五种主要数据最小化策略：

**终止对话 (P4)**。人们选择结束对话以避免进一步信息共享。“我并未改变已设定的提示，但确实停止了对话”，这表明一种在对话开始后避免个人暴露升级的策略。

**避免记录 (P3、P4、P9)**。许多参与者警惕在ChatGPT中留下永久记录，主要担忧其作为基于大型语言模型的系统的可学习性与记忆功能。P9解释他们会关闭记忆功能，并表示：

“我正在旅行...家里有棵植物...叶子变黄了...所有这些都被记录在记忆功能里，这绝对不是我想要的。”P4会删除聊天记录以消除痕迹，突显了避免留下可追踪数字足迹的意愿。

**省略细节 (P1、P8、P11)**。部分参与者有意省略关键个人信息，如姓名或财务信息。P11解释道：“我只是没有提供我的姓名。”类似地，P1出于商业目的避免包含财务详情，“如果我要用它来发展业务，我不会输入.....个人信息。”这些示例表明参与者刻意限制所共享数据的敏感性。

**伪造数据 (P4、P5、P9)**。部分参与者提供虚假信息以掩盖其身份或情况，通常会改变故事主题或查询内容，避免信息直接关联到自身。例如，P5和P9对数据可追踪性表示担忧，促使他们更改姓名。此外，

参与者即使在未亲身涉及的情况下也会模糊细节。P4举例说明：“我有个员工刚生了孩子.....我选择用朋友而非员工来表述...这显得更谨慎。”概括信息 (P2、P8)。另一种常见策略是通过概括查询来避免分享精确或可识别细节。例如P8指出，在提交文件供审查时会修改或摘要内容：“我会稍作编辑或写摘要.....不包含超级重要、更保密[信息]。”

参与者们承认，即便采取了这些策略，一旦数据被共享，他们也无法完全掌控。P9指出，尽管他们试图避免某些操作，但系统的

memory might still retain their information. Similarly, P4 articulated a fear of losing control, believing the data may never truly be erased once shared.

**6.1.5 Continuous usage and trade-off (P3, P4, P9).** Despite concerns, all participants continued to use ChatGPT due to its convenience. As P4 mentioned, they were willing to “risk the privacy” to “save the time”; while P9, despite worries about exposing financial data, used GPT for financial advice. P3 captured the “inevitable” trade-off, stating, “I think it’s kind of a dilemma right? Because I do try to avoid giving more information than necessary, but it feels... difficult to get what I want..., without providing... information inevitably” (P3). These struggles highlight a clear demand for privacy-preserving data minimization support in LLM-based conversational agents.

## 6.2 Factors for Privacy Trust

We identified six factors impacting user privacy trust over interaction with ChatGPT and Rescriber to evaluate and inform the design of a trusted, user-led data minimization tool.

**6.2.1 Transparency and On-Device Needs of the Extension (P4, P7).** A key factor influencing user privacy trust is the need for transparency and an on-device implementation of the data minimization tool, which aligns with our goal of leveraging smaller models hosted locally (e.g., Rescriber-Llama3-8B). Participants stressed the importance of understanding how their data is managed. “that’s just a matter of transparency... you never know that they’re not [collecting information].” (P7) Participants expected local processing to avoid replicating their concerns with ChatGPT. P4 emphasized, “I assume that the extension is running itself locally and is not storing any of the data...” (P4), reflecting users’ need for control over their data.

**6.2.2 Consistent Detection Results (P4, P9, P11).** A critical factor affecting user trust in Rescriber was the consistency of detection results. Participants expressed frustration when the similar inputs produced different feedback, which undermined their confidence in the tool’s reliability. P11 noted confusion when one model detected fewer privacy risks than the other, whereas P9 observed that Rescriber missed sensitive information when given partial inputs, while caught them when given the full messages. Similarly, P4 expressed skepticism when Rescriber-GPT-4o flagged sensitive content during a second run not the first: “it loses... credibility... making me skeptical of robustness of the tool.” (P4) This lack of consistency, sometimes due to the non-deterministic nature of the model, led participants to question the effectiveness of Rescriber’s capabilities.

**6.2.3 Consistent Abstracting Results (P5, P10).** Users’ confidence in Rescriber is closely tied to the consistency of its abstraction results. Inconsistent abstraction can lead to doubts about the system’s ability to protect sensitive data. P5 expressed concern when Rescriber-GPT-4o displayed inconsistency in abstracting “KCMO” also due to its non-deterministic nature, first changing it to “major city in Midwest” and later to “my city,” prompting P5 to question the integrity of the responses. In contrast, P10 reported satisfaction with Rescriber-Llama3-8B for its consistent abstraction, noting, “I’m satisfied because the same information is again over here”.

**6.2.4 Completeness of Detection (P2, P4, P5, P6, P8).** A key factor influencing user trust in Rescriber is the completeness of its detection and sanitization. Participants felt more confident and protected when the system detected a broader range of sensitive data. P5 highlighted the tool’s ability to catch overlooked details, “it identified things that I didn’t even pick up could be sensitive”.

Participants often compared models to assess detection. P5 preferred Rescriber-Llama3-8B which detected more identifiers like height and weight, while P2 found Rescriber-GPT-4o more thorough, “we picked up way more on this one... so this one is doing much better.” These examples are related to the advantages of the two models in different aspects. Rescriber-GPT-4o flagged more entities overall, as echoed by our quantitative findings (subsection 5.1); and Rescriber-Llama3-8B excelled at detecting entities outside of the specified taxonomy, possibly due to its weaker instruction-following capabilities which made it more likely to not be constrained by the taxonomy.

Participants’ faith in the system diminished with incomplete detection. P6 expressed doubt when they “don’t necessarily know that it caught all the personal information.” The consequence could be catastrophic. P4 encountered a case where the model failed to detect a piece of sensitive information he expected it to catch, completely disrupting his trust. He remarked, “It loses all of its reputability with me right away.” These examples reinforce the importance of thorough detection for fostering user trust in privacy-preserving tools like Rescriber.

**6.2.5 Completeness of Sanitization (P3, P4, P6, P7).** Participants emphasized the importance of providing sanitization features to ensure complete data minimization. Rescriber’s ability to aggregate all instances of the same data type was praised by P4, stating, “I really enjoyed that everything was sorted properly... I’m not missing one because they’re all there.” The “select all” option was also valued for simplifying complete privacy protection, as put by P7, “that would be much easier if you would do that in one click.”

Participants sometimes even expressed a desire for more assertive privacy-preserving mechanisms. P6 remarked on the system’s limitations in enforcing data minimization: “it helps me reduce it, but it doesn’t necessarily force me to do that,” (P6) reflecting a wish for a more proactive approach to completely redact sensitive data.

**6.2.6 Control over Sanitization (P4, P5, P6, P7).** Another factor in building user trust in Rescriber is giving participants control over the sanitization process. Participants reported greater trust and satisfaction given the autonomy to decide which data to mask. “I have more control over the information I’m sharing.” (P6). P5 further highlighted the empowerment from this control, explaining, stating, “it made me aware of those different areas and gave me... the final choice.” (P5) This sense of control allowed users to feel in charge of their privacy decisions.

## 6.3 Factors for Adoption

We identified key factors influencing user adoption of Rescriber for privacy-preserving data minimization practices.

记忆功能可能仍会保留他们的信息。类似地, P4表达了对失去控制的担忧, 认为数据一旦共享可能永远无法真正被抹除。

**6.1.5 持续使用与权衡 (P3, P4, P9)。** 尽管存在顾虑, 所有参与者仍因ChatGPT的便利性而继续使用它。如P4所述, 他们愿意“冒险牺牲隐私”以“节省时间”; 而P9尽管担心暴露财务数据,

仍使用GPT获取财务建议。P3道出了这种“不可避免”的权衡: “我觉得这有点像两难对吧? 因为我确实尽量避免提供不必要的信息, 但感觉.....不提供.....信息就难以得到我想要的.....” (P3)。

这些困境凸显了在基于LLM的对话代理中对隐私保护数据最小化支持的明确需求。

## 6.2 隐私信任的影响因素

我们确定了六个影响用户在与ChatGPT和Rescriber互动过程中隐私信任的因素, 以评估并指导可信赖、用户主导的数据最小化工具的设计。

### 6.2.1 扩展的透明度与设备端需求 (P4, P7)。

影响用户隐私信任的一个关键因素是对数据最小化工具的透明度需求及其设备端实现, 这与我们利用本地托管的小型模型(如Rescriber-Llama3-8B)的目标一致。参与者强调了理解其数据如何被管理的重要性。“这只是透明度的问题.....你永远不知道他们是否没有[收集信息]。” (P7) 参与者期望通过本地处理来避免重蹈他们对ChatGPT的担忧。P4强调说:

“我假设扩展是在本地自行运行, 并未存储任何数据.....” (P4), 这反映了用户对数据控制的需求。

### 6.2.2 一致的检测结果 (P4, P9, P11)。

影响用户对Rescriber信任的关键因素是检测结果的一致性。参与者对相似输入产生不同反馈表示沮丧, 这削弱了他们对工具可靠性的信心。P11指出, 当一个模型检测到的隐私风险少于另一个时感到困惑, 而P9观察到Rescriber在输入不完整时会遗漏敏感信息,

但在输入完整消息时却能捕捉到这些信息。类似地, P4对Rescriber-GPT-4o在第二次而非首次运行时标记敏感内容表示怀疑: “这降低了.....可信度.....”

让我对工具的稳健性产生怀疑。” (P4) 这种不一致性有时源于模型的非确定性本质, 导致参与者质疑Rescriber能力的有效性。

### 6.2.3 一致的抽象化结果 (P5, P10)。

用户的信心与Rescriber抽象化的一致性密切相关。不一致的抽象化可能导致对系统保护敏感数据能力的质疑。P1表示担忧当Rescriber-GPT-4o在抽象化“KCMO”时表现出不一致性同样由于其非确定性, 首次将其改为

“中西部主要城市”, 随后又改为“我的城市”, 这促使P5质疑响应的完整性。相比之下, P10报告对Rescriber-Llama3-8B模型因其一致性感到满意

抽象化, 并指出, “我很满意, 因为同样的信息是再次出现在这里”。

**6.2.4 检测完整性 (P2, P4, P5, P6, P8)。** 影响用户对Rescriber信任的关键因素是其检测与清理的全面性。当系统能检测到更广泛的敏感数据时, 参与者会感到更安心且受保护。P5特别强调了该工具捕捉被忽视细节的能力, “它识别出了我甚至没意识到可能是敏感信息的内容”。

参与者经常比较模型以评估检测效果。P5更倾向于Rescriber-Llama3-8B模型, 该模型检测到了更多标识符, 如身高和体重, 而P2认为Rescriber-GPT-4o更为全面, “我们在一个上发现了更多...所以这一个做得更好。”这些示例与两种模型在不同方面的优势相关。Rescriber-GPT-4o总体上标记了更多实体, 这与我们的定量研究结果相呼应(见第5.1小节); 而Rescriber-Llama3-8B在检测指定分类法之外的实体方面表现优异, 可能是由于其较弱的指令遵循能力使其更容易受到分类法的限制。

参与者对系统的信任因不完全检测而减弱。P6表达了疑虑, 当他们“不一定知道它捕捉到了所有的个人信息”时。后果可能是灾难性的。P4遇到了一个案例, 模型未能检测到他预期会捕捉到的一条敏感信息,

彻底破坏了他的信任。他评论道: “它立刻在我眼中失去了所有可信度。”这些示例强化了全面检测对于培养用户对Rescriber等隐私保护工具信任的重要性。

**6.2.5 清理的完整性 (P3, P4, P6, P7)。** 参与者强调了提供清理功能以确保数据最小化完整性的重要性。P4赞扬了Rescriber能够聚合相同类型的所有数据实例: “我真的很喜欢一切都分类得当...我不会漏掉任何一个, 因为它们都在那里。”

“全选”选项也因简化完整的隐私保护而受到重视, 正如P7所说: “如果能一键完成, 那会容易得多。”

参与者有时甚至表达了对更积极的隐私保护机制的渴望。P6评论了系统在强制执行数据最小化方面的限制: “它帮助我减少数据, 但并未强制我这样做,” (P6) 反映出对更主动彻底编辑敏感数据的愿望。

**6.2.6 清理控制 (P4, P5, P6, P7)。** 建立用户对Rescriber信任的另一个因素是让参与者能够控制清理过程。参与者表示, 由于拥有自主决定哪些数据需要屏蔽的权利, 他们的信任感和满意度更高。

“我对共享的信息有了更多控制权。” (P6)。P5进一步强调了这种控制带来的赋权感, 并解释道:

“它让我意识到这些不同的领域, 并给了我.....最终的选择权。” (P5) 这种控制感让用户感到自己能够主导隐私决策。

## 6.3 采用因素

我们识别了影响用户采用Rescriber的关键因素: 隐私保护的数据最小化实践。

**6.3.1 Balancing Privacy with Task Effectiveness (All but P4, P7, P11).** Rescriber's ability to generalize sensitive information without compromising task effectiveness influenced user adoption. Participants were satisfied when abstractions preserved contexts. For instance, P3 approved of abstracting "Yonkers, NY" to "a city in NY," as this generalization still allowed them to find recommended places within the region. P10 valued abstracting "John Doe" with "a traveler" in E2, which kept the travel data analysis task intact. Nevertheless, when abstraction distorted the original intent, participants were less likely to adopt the system's solutions. P6 rejected the abstraction of "downtown Troy" to "State Park," explaining that the context had shifted too far from their original intent.

Participants were not always able to predict the information that optimized privacy-utility trade-offs. P2 removed some health information from her prompt as suggested by Rescriber, expecting less thorough feedback. However, after comparing the new response based on the redacted prompt with the original, she found them very similar, and said, "*I think I'm satisfied... it doesn't seem to be as detrimental as I thought.*"

**6.3.2 Convenience (P3, P7, P8, P10, P11).** Overall, participants valued the convenience provided by Rescriber, recognizing its ability to simplify complex privacy-preserving tasks, which drives user adoption. P7 emphasized the importance of simplicity, noting it, "*simplify the whole process of using something that's already supposed to be easy.*"

Another key convenience factor was the available features that streamline data sanitization:

- **Select all:** P8 appreciated the "select all" feature, saying, "*convenience and just helping you get stuff done at a faster rate.*"
- **Replace back:** Participants noted that replacing sanitized terms back into the conversation improved readability. P3 remarked on its usefulness for comprehension, stating, "*really useful for keeping comprehension up, because if we just gave the default names, you'll get really confusing.*"

**6.3.3 The tool should respect people's habit of using ChatGPT (P4, P6, P8, P11).** Another factor influencing user adoption of Rescriber is how the tool affects their regular interaction patterns with ChatGPT. Participants expressed a reluctance to change their established workflow, which prioritizes the speed and fluidity of exchanges with the chatbot. For instance, P4 described their typical behavior when using ChatGPT as a back-and-forth interaction without much interruption, "*It's a shame that I would have to change the way that I operate with ChatGPT, because my instinct is to treat it like a chat bot, just input...type [enter to send].. input.. type [enter to send].. input...*" (P4) This highlights a friction between the participants' habitual use of ChatGPT and the interruption caused by privacy protection measures.

**6.3.4 Having Alternative Solutions as a Benefit (P3, P5).** One final key adoption factor was the availability of actionable solutions. Participants appreciated that Rescriber not only identify sensitive information but also offer practical alternatives to sanitize it. P3 highlighted the challenge of manually abstracting sensitive information: "*having something that can do it for you is really useful.*" P5 similarly valued the tools providing actionable solutions, rather

than just flagging sensitive content: "*provides solutions instead of just marking a word.*"

## 6.4 Data Minimization Strategies in Reaction to the Tools

We uncovered nuanced behaviors and decision-making processes that participants adopted to minimize the data they shared with ChatGPT during testing. Their decisions were influenced by the task context and sensitivity of the data.

**6.4.1 What to sanitize: balancing sensitivity and task relevance.** Participants tailored data sanitization based on whether masking certain information would impact the quality of the response. They aimed to obscure personal or sensitive data without compromising the model's output, but opinions differed on what constituted sensitive information.

**Task relevance (P6, P11).** Some participants removed unnecessary details that were perceived to not affect the task, like P6 who noted that details about why someone visited a treatment center weren't needed for writing a letter. Conversely, P11, argued that certain details could not be obscured without negatively impacting the conversation's purpose, particularly, in medical contexts where specifics are crucial. **Removing partial information as a compromise (P2, P4, P5, P6, P11).** Many participants preferred to obscure only part of the identifiable data, believing that hiding one or two key details was enough to preserve privacy and get useful feedback. P5 felt that altering height or weight alone could suffice for maintaining privacy while not affecting the task, explaining that, "*changing one of them... would be sufficient.*" (P5)

**Individual differences (P2, P5, P7).** We also observed individual differences on data minimization strategies over a same case, E3, about writing a thank-you note to the hypothetical user's therapist. For example, P5 was comfortable obscuring some health details when writing a thank-you note to a therapist, while P2, worried that over-sanitizing health-related information would lead to overly generic or less useful responses: "*If we're keeping this stuff out of GPT, it's not going to be able to provide as thorough of an answer.*" (P2) P7 emphasized that for domain-specific tasks (e.g., legal questions) people may have different levels of expectations of the specificity of the answer based on their domain knowledge, explaining, "*it does help protect your privacy if you don't know much about the law and you're just getting a baseline...it depends on the knowledge level of the person using it.*"

**Name data as a high-priority concern (P1, P3, P4).** Across multiple participants, names emerged as a particularly sensitive data type, with many users prefer to remove or anonymize them, echoing the quantitative analysis result which showed name as the most frequently redacted or abstracted data type (subsection 5.1). P3 noted that names were non-essential for many cases and should be removed first, which reflects a common concern across participants.

**6.4.2 Rationales for choosing replacement or abstraction.** Participants adopted different strategies to choose between replacing data with placeholders or abstracting the information, depending on the context and their privacy concerns.

**Abstraction for retaining context (P3, P5, P7, P8).** Abstraction was favored for conveying general contextual information

仅仅标记敏感内容: "它提供解决方案, 而不仅仅是标出一个词。"

## 6.4 针对工具的数据最小化策略 工具

我们揭示了参与者为最小化与ChatGPT测试期间共享的数据所采取的微妙行为和决策过程。他们的决策受到任务背景和数据敏感性的影响。任务背景和数据敏感性。

**6.4.1 清理内容: 平衡敏感性与任务相关性。** 参与者根据遮蔽特定信息是否会影响响应质量来定制数据清理策略。他们的目标是隐藏个人或敏感数据而不损害模型输出, 但对于何为敏感信息的界定存在分歧。

任务相关性 (P6、P11)。部分参与者会移除被认为不影响任务的多余细节, 例如P6指出写信时无需说明某人为何前往治疗中心的具体原因。相反, P11主张某些细节若被遮蔽会损害对话目的—尤其在医疗情境中, 具体信息至关重要。折衷方案: 部分信息清理 (P2、P4、P5、P6、P11)。多数参与者倾向于仅遮蔽部分可识别数据, 认为隐藏一两个关键细节既能保护隐私又能获得有用反馈。P5认为仅修改身高或体重就足以维护隐私且不影响任务, 并解释称: "改动其中一项……就足够了。" (P5)

个体差异 (P2、P5、P7)。我们还观察到, 针对同一案例E3, 不同参与者对数据最小化策略的实施存在个体差异,

关于为假设用户向其治疗师撰写感谢信的场景。示例: P5在给治疗师写感谢信时愿意模糊化某些健康详情, 而P2则担心过度清理健康相关信息会导致响应过于笼统或缺乏实用性: "如果我们把这些内容从GPT中剔除, 它就无法提供足够详尽的答案。" (P2) P7强调, 对于特定领域任务 (如法律问题), 人们基于自身领域知识可能对答案的详细程度有不同预期, 并解释道: "如果你对法律了解不多, 只是获取基础信息时, 这确实有助于保护隐私……这取决于使用者的知识水平。"

姓名数据作为高优先级关注点 (P1、P3、P4)。多位参与者均指出姓名属于特别敏感的敏感数据类型, 许多用户倾向于删除或匿名化处理, 这与定量分析结果相呼应, 该结果显示姓名为最常被删除或抽象化的数据类型 (第5.1小节)。P3指出, 在许多情况下姓名并非必要信息, 应优先移除, 这反映了参与者们普遍的隐私顾虑。

**6.4.2 选择替换或抽象的理据。** 参与者们根据上下文及其隐私顾虑, 采用了不同策略来选择是用占位符替换数据还是将信息抽象化。

保留上下文的抽象化 (P3、P5、P7、P8)。抽象化被推荐用于传达一般上下文信息

**6.3.1 隐私与任务效果的平衡 (除P4、P7、P11外)。** Rescriber在不影响任务效果的前提下泛化敏感信息的能力影响了用户采用率。当抽象化保留了上下文时, 参与者表示满意。例如,

P3赞同将“纽约州扬克斯”抽象化为“纽约州的一个城市”, 因为这种泛化仍能让他们找到该区域的推荐地点。P10在E2中重视将“约翰·多伊”泛化为“一位旅行者”, 这保持了旅行数据分析任务的完整性。然而, 当抽象化扭曲了原始意图时, 参与者不太可能采纳系统的解决方案。P6拒绝将“特洛伊市中心”抽象化为“州立公园”, 解释说上下文已偏离其原始意图太远。

参与者并不总能预测出优化隐私与效用权衡的信息。P2删除了部分健康根据Rescriber的建议, 从她的提示中获取信息, 期望得到不那么详尽的反馈。然而, 在将基于编辑后提示的新响应与原始提示的响应进行比较后, 她发现两者  
非常相似, 并表示: “我想, 我满意了……似乎并不像我原先以为的那样有害。”

**6.3.2 便利性 (P3、P7、P8、P10、P11)。** 总体而言, 参与者重视Rescriber提供的便利性, 认可其简化复杂隐私保护任务的能力, 这推动了用户采用。P7强调了简洁性的重要性, 指出, “简化本应简单之物的整个使用流程。”

另一个关键的便利性因素是那些简化数据清理的可用功能:

- **全选:** P8赞赏“全选”功能, 表示, “便利性且能帮助你更快完成任务。”
- **替换回:** 参与者指出, 将清理过的术语替换回对话中提升了可读性。P3评论其对于理解的有用性, 称“确实有助于保持理解力, 因为如果我们只提供默认名称, 会让人非常困惑。”

**6.3.3 该工具应尊重人们使用ChatGPT的习惯 (P4, P6, P8, P11)。** 影响用户采用Rescriber的另一个因素是该工具如何影响他们与ChatGPT的常规交互模式。

参与者表示不愿改变其既定工作流程, 该流程优先考虑与聊天机器人交流的速度和流畅性。例如, P4描述了他们在使用ChatGPT时的典型行为, 即一种几乎没有中断的来回互动, “很遗憾我不得不改变与ChatGPT的交互方式, 因为我的本能是将其视为一个聊天机器人, 只需输入...键入[回车发送]..输入..键入[回车发送]...输入...” (P4) 这凸显了参与者对ChatGPT的习惯性使用与隐私保护措施造成的干扰之间的摩擦。

**6.3.4 提供替代解决方案作为优势 (P3、P5)。** 最后一个关键的采用因素是可操作解决方案的可用性。

参与者们赞赏Rescriber不仅能识别敏感信息, 还能提供实用的替代方案来清理这些信息。P3强调了手动提取敏感信息的挑战: “有工具能替你完成这件事真的很有用。” P5同样看重该工具提供的可操作的解决方案, 而非

without revealing specifics. For example, P3 opted for abstracting numbers to maintain content clarity without divulging precise information, stating that abstraction helps “*get an idea without providing specific numbers.*”

**Replacement for non-essential data (P3, P12).** Conversely, when participants encountered data irrelevant to the task, they preferred to use direct replacements. Names were frequently cited as data that could be safely replaced without affecting task outcomes. “*I'm gonna go for the names... I don't think that's relevant to this question.*” (P3) These strategies underscore how participants actively engaged with the tool to tailor their privacy measures based on the specific task, data type, and expected response quality.

## 6.5 Education and Learning Process

During the testing, we observed that the extension played an educational role, shaping participants' mental models of data privacy and increasing their awareness of sensitive information.

**6.5.1 Educational Value of the Tool.** We summarize the educational values perceived by our participants below.

**Learning from the detection results (P2, P3, P4, P5, P6, P7, P11).** The tool reminded users to reassess the data they share, as P7 mentioned, “*the idea of highlighting it is a helpful thing. It lets you spot things that you may have put in there by accident, or the you may have just glossed over.*” Additionally, the tool helped participants recognize information they might not have considered sensitive. For example, P3 was surprised when “Barnes and Nobles” (the name of a book store) was flagged, but acknowledged it made sense when considering that there are not many of them in their location, which was specified in the prompt. This highlights how the tool prompts users to identify unrecognized privacy risks.

**Impact beyond one-time education (P2, P4, P5, P9, P11).** The tool's educational impact extended beyond one-time interactions. Participants' understanding of sensitive data evolved with use. For instance, P5 experienced a shift in mental models after testing the extension across multiple sessions. They learned to mask certain terms like “polyvagal therapist” as they saw previous models flagged it and even when it was not flagged by a later model, suggesting the tool fosters awareness for safeguarding personal information. Even brief exposure to Rescriber enhanced participants' ability to recognize sensitive data, indicating a lasting educational effect.

**6.5.2 Learning process of how to do sanitization.** Participants faced challenges with the abstraction feature, which was less intuitive at first. P2 admitted forgetting about abstraction, while P8 confused it with the replacement feature, expecting it to “*change it to more general terms.*” Additionally, people often lacked clear expectations of abstraction results, while they like to see things match their expectation.

Participants also experienced a learning curve in understanding the balance between safeguarding sensitive information and maintaining the quality of responses generated by Rescriber. P2 doubted the system's ability to generate thorough answers after redacting health information. However, after testing, they found the sanitized input yielded similar-quality responses, realizing “*the substance [of the response before and after sanitization] isn't very*

*different.*” The evolution in participants' assessments suggests that while initial hesitation exists, confidence grows as users see that redaction doesn't significantly impact response quality.

## 7 Discussion

We present discussions of our results, including the proposal of smaller-LLM-based privacy controls, validated by our studies as a promising solution to privacy issues in LLM-based systems; a summary of the value of user-led data minimization; and synthesized design implications for AI-powered, human-in-the-loop privacy controls.

### 7.1 On the Feasibility of Smaller-LLM-Based Privacy Controls of LLM-Based Systems

Smaller LLMs (e.g., Llama3-8B) are widely considered to have the benefit of inherent privacy protection, as they can be deployed locally to avoid concerns associated with sharing data with larger LLMs' service providers. Research and real-world applications have primarily focused on fine-tuning these smaller LLMs for specific downstream tasks (e.g., proofreading) to enhance their performance on those tasks. In this work, we propose a different approach, leveraging smaller LLMs to empower privacy controls, particularly by helping users minimize unnecessary disclosure when interacting with LLM-based conversational agents. Our prototyping and evaluation results have demonstrated the feasibility and potential of this idea in multiple aspects.

First, **Rescriber-GPT-4o** and **Rescriber-Llama3-8B** achieved similar subjective perceptions regarding data minimization efficacy, reduced data sharing, privacy protection, and adoption intention (subsection 5.1). Participants rated both versions of the tool favorably. In our qualitative analysis (subsection 6.2), we attributed the varied preferences to different areas that the two versions of the system do well in.

We also observed a disparity between the subjective ratings, the objective reduction, and the accuracy benchmarking of the two models. Although participants liked both versions, the actual reduction of disclosure when using **Rescriber-Llama3-8B** is half of that of **Rescriber-GPT-4o** (subsection 5.1). This observation corroborates the benchmarking of PII detection based on our taxonomy which reveals that GPT-4o outperforms Llama3-8B in recall (subsubsection 3.5.3). This disparity is fundamentally related to users' limited experiences and established standards in the task. We observed that users did not evaluate the tool based on a comprehensive set of predefined ground truths for detection. Instead, they were often prompted by our tool to become aware of sensitive disclosures that they had not realized before (subsubsection 6.5.1). In this aspect, both models did a good job and significantly improved upon the status quo, as put by P5 when asked to compare the two versions, “*if I'm answering the question in a vacuum, and just comparing it against what ChatGPT originally provides, then it's a massive step up.*”

The disparity between subjective ratings and objective reductions raises questions about how to prevent users from having a *false sense of privacy*. Specifically, how to address their over-reliance on the tool, with the assumption that it will completely remove their personal information, even though it may actually miss some. One

同时避免透露具体细节。例如, P3选择对数字进行抽象化, 以在不泄露精确信息的情况下保持内容清晰度, 并表示抽象化有助于“了解大致情况而不提供具体数字”。

**对非必要数据的替换 (P3、P12)。**相反, 当参与者遇到与任务无关的数据时, 他们更倾向于使用直接替换。姓名常被提及为可安全替换且不影响任务结果的数据。

“我打算替换姓名……我认为这与问题无关。”(P3)这些策略突显了参与者如何积极利用工具, 根据具体任务、数据类型及预期响应质量定制隐私措施。

### 6.5 教育与学习过程

在测试过程中, 我们观察到该扩展发挥了教育作用, 塑造了参与者对数据隐私的心智模型, 并增强了他们对敏感信息的意识。

**6.5.1 工具的教育价值。**我们总结了参与者感知到的教育价值如下。

**从检测结果中学习 (P2、P3、P4、P5、P6、P7、P11)。**该工具提醒用户重新评估他们分享的数据, 正如P7所言, “高亮显示的想法很有帮助。它能让你发现可能无意中放入的内容, 或者你可能只是略过的内容。”此外, 该工具帮助参与者认识到他们可能未曾视为敏感的信息。例如, P3惊讶地发现“巴诺书店”

(书店名称)被标记了, 但考虑到提示中指定的所在位置这类书店并不多, 参与者认为这合乎情理。这凸显了该工具如何提示用户识别未被察觉的隐私风险。

**超越一次性教育的影响 (P2、P4、P5、P9、P11)。**该工具的教育效果不仅限于单次交互。参与者对敏感数据的理解随着使用而深化。例如, P5在多次测试扩展功能后心智模型发生了转变。他们学会屏蔽“多迷走神经治疗师”等特定术语, 因为发现早期模型会标记这些内容, 即使后续模型未作标记,

这表明该工具能有效提升保护个人信息的意识。即使短暂接触Rescriber, 也增强了参与者识别敏感数据的能力, 显示出持久的教育效果。

**6.5.2 数据清理的学习过程。**参与者面临抽象功能带来的挑战, 起初并不直观。

P2承认忘记了抽象化, 而P8则将其与替换功能混淆, 期望它能“*将其改为更通用的术语*”。此外, 人们往往对抽象结果缺乏明确的预期, 而他们更希望看到事物符合自己的抽象结果中, 他们喜欢看到事物与其相匹配期望。

参与者在理解保护敏感信息与保持Rescriber生成响应质量之间的平衡上也经历了一段学习曲线。P2曾怀疑该系统在编辑健康信息后生成详尽答案的能力。然而经过测试, 他们发现清理后的输入能产生质量相似的回应, 并意识到“*清理前后响应内容[的实质]差异不大*”。

参与者评估的演变表明尽管最初存在犹豫, 但随着用户看到编辑不会显著影响响应质量, 信心逐渐增强。

## 7 讨论

我们围绕研究结果展开讨论, 包括提出基于小型LLM的隐私控制方案——该方案经验证能有效解决LLM系统中的隐私问题; 总结用户主导数据最小化的价值; 并综合设计启示, 为人工智能驱动的、人在回路的隐私控制系统提供参考。

### 7.1 基于小型LLM的可行性探讨

#### LLM系统中的隐私控制

小型大语言模型(如Llama3-8B模型)被普遍认为具有内在隐私保护的优势, 因为它们可以本地部署, 从而避免与大型语言模型服务提供商共享数据相关的担忧。研究和实际应用主要集中于对这些小型大语言模型进行特定下游任务(如校对)的微调, 以提升其在这些任务上的性能。在本研究中, 我们提出了一种不同的方法, 利用小型大语言模型来增强隐私控制, 特别是通过帮助用户在与基于LLM的对话代理互动时尽量减少不必要的披露。我们的原型设计和评估结果从多个方面证明了这一想法的可行性和潜力。

首先, **Rescriber-GPT-4o**和**Rescriber-Llama3-8B**在数据最小化效能、减少数据共享、隐私保护及采用意向上获得了相似的主观感知评价(第5.1小节)。参与者对两个版本的工具均给予积极评价。在我们的定性分析中(第6.2小节), 我们将这种偏好差异归因于系统两个版本在不同领域的出色表现。

我们还观察到两个模型在主观评分、客观减少量和准确性基准测试之间存在差异。尽管参与者喜欢两个版本, 但实际使用**Rescriber-Llama3-8B**时的信息披露减少量仅为**Rescriber-GPT-4o**的一半(见第5.1小节)。这一观察结果与我们基于分类法的个人身份信息检测基准测试相吻合, 该测试显示GPT-4o在召回率上优于Llama3-8B模型(见第3.5.3小节)。这种差异从根本上与用户在该任务中有限的经验和既定标准有关。

我们注意到用户并未基于一套预定的检测基准真相来全面评估该工具, 而是经常被我们的工具提示注意到之前未曾意识到的敏感信息披露(见第6.5.1小节)。

在这方面, 两个模型都表现出色, 并显著改善了现状。正如P5在被要求比较两个版本时所说:“如果我在真空环境中回答问题, 仅将其与ChatGPT最初提供的答案相比, 那这是一个巨大的进步。”

主观评分与客观降低之间的差异引发了如何防止用户产生虚假的隐私感的问题。具体而言, 如何解决他们对工具的过度依赖, 即假设该工具会完全移除他们的个人信息, 而实际上可能会遗漏部分信息。

direction is to keep improving the accuracy of the small model for detection by fine-tuning the models or combining the results of multiple models. Another direction is to calibrate users' expectations of the capability and scope of detection with the actual model's capability. For example, we noticed that participants spontaneously compared the suggestions of the two models. When they encountered different results between the two models, they realized that one model did a better job than the other, and certain information was missing. We may intentionally display results from multiple models to users to expose them to the potential risks of missed detection and reduce overreliance.

In addition to improvements to the small models, the response generation also needs enhancements to better accommodate the use of placeholders to generate accurate and correctly formatted responses, as indicated in our response utility analysis (see subsection 5.2). Future work could explore introducing a stricter control mechanism during response generation. For example, explicitly instructing the model to avoid creating new placeholders or to follow a predefined mapping schema could be beneficial. Additionally, using in-context learning to provide examples might prevent overly conservative answers when encountering placeholders.

Overall, these findings suggest that smaller LLMs, even without fine-tuning, already possess privacy-preserving capabilities to an extent that can be converted into enhanced privacy protection to end users. However, room for improvement is also clear, particularly in terms of metrics that affect the user-led data minimization results.

## 7.2 User-Led Data Minimization Bridges Gaps in Privacy Awareness and Controls

Our study examines people's privacy concerns about using Chat-GPT, focusing on the information identifiability and the lack of transparency. Participants employed some data minimization strategies in their everyday usage, but found them inadequate, leaving them feeling vulnerable and lacking control (subsection 6.1). These findings validate the general need to offer users convenient and flexible support for data minimization and also generate deeper insights into how certain design choices of Rescriber are useful to handle the nuanced and personal user risk perceptions. For example, giving users full control over the sanitization actions was appreciated by users (subsubsection 6.2.6), and also helps them manage the privacy-utility tradeoffs following their preferences (e.g., varied assessment of identifiability in the same case, see subsection 6.1.1 and subsubsection 6.4.1). Users valued our tool and were willing to adopt it especially because of many features we provided to streamline the process, such as providing alternative solutions (subsubsection 6.3.4), convenience features like "Select all", and "Write back" (subsubsection 6.3.2), and helping users find the middleground between privacy and utility by offering two options of sanitization targeting different scenarios, abstraction and replacement (subsubsection 6.4.2).

Another gap our tool addresses is increasing users' awareness of sensitive disclosure (subsubsection 6.5.1). Participants frequently cited the tool's educational value as a benefit alongside enhanced privacy control and protection. Although not specifically designed as an educational tool, our tool demonstrates an example of achieving privacy education by providing in-situ learning opportunities

about the privacy risks in regular data-sharing, similar to the idea of experiential learning [16]. Users were willing to engage in exploring different sanitization options because it provided immediate and tangible benefits to them. This insight could inform more privacy education design and research situated in the context of user-led data minimization.

## 7.3 Implications for Designing AI-Powered Human-in-the-Loop Privacy Controls

Our approach highlights the need for a shift from traditional human-out-of-the-loop Privacy-Enhancing Technologies (PETs) to tools that support human-in-the-loop, granular privacy management. While AI is useful in achieving this goal, properly integrating AI into such systems, so that users will trust it, be willing to use it, and find that the protection it offers aligns with their expectations, requires an in-depth understanding of the dynamics of interactions between humans and AI, as moderated by the interfaces between them. We reflect on lessons learned from our studies on designing such systems and interfaces below.

**Towards human-centered evaluation.** Although performance metrics like accuracy, precision, and recall are all important ways to assess the performance of a model, their impact on user experience and decision-making varies significantly. As highlighted in subsection 6.2, the comprehensiveness and consistency of the detection and sanitization results are key factors that affect users' trust. Specifically, users were amenable towards false positives in detection results (e.g., less sensitive information) and mislabeled data categories, whereas a single false negative could be catastrophic to their trust in the system. This suggests that certain metrics, like recall, are more important than others, like precision.

Additionally, we found that users' perceptions of sensitive disclosure span a wide range of topics, which are beyond the scope of our taxonomy and likely cannot be fully captured by any taxonomy (subsubsection 6.2.4). When Rescriber-Llama3-8B generated detection results outside the taxonomy (e.g. weight and height), users actually preferred them over Rescriber-GPT-4o, which more strictly adheres to the taxonomy and did not detect them as sensitive data. However, common benchmarking methods driven by a taxonomy will misrepresent users' preference of this case in the result, which indicates the limitations of the taxonomy-based evaluation method and emphasizes the importance of privacy protection beyond PII [6].

**Balance between comprehensiveness and control.** In subsubsection 6.2.6 and subsubsection 6.2.5, we observed seemingly conflicting needs concerning privacy controls. We also saw in subsubsection 6.3.1 that users struggle to identify the optimal trade-offs between privacy and utility, due to a lack of a deep insights of the models' capabilities. This tension highlights the need to strike a balance between giving users control to establish subjective trust and ensuring comprehensive data minimization without being undermined by users' limited experiences or attention spans. Future research can further explore different levels of user engagement and agency, and the impact of individual difference, to design better interfaces that meet both important requirements.

一个方向是通过微调模型或结合多个模型的结果，持续提升小型检测模型的准确性。另一个方向是根据模型的实际能力，校准用户对检测能力和范围的期望。例如，我们注意到参与者会自发比较两个模型的建议。当他们遇到两个模型结果不一致时，意识到其中一个模型的表现优于另一个，且某些信息被遗漏。我们可以有意向用户展示多个模型的结果，使其意识到漏检的潜在风险，从而减少过度依赖。

除了对小型模型的改进外，响应生成也需要增强以更好地适应使用占位符生成准确且格式正确的响应，如我们的响应效用分析所示（参见第5.2小节）。未来的工作可以探索在响应生成过程中引入更严格的控制机制。例如，明确指示模型避免创建新的占位符或遵循预定义映射模式可能是有益的。此外，

使用上下文学习提供示例可能会在遇到占位符时防止过于保守的答案。

总体而言，这些发现表明，小型大语言模型即使未经微调，也已具备一定程度的隐私保护能力，可以转化为对终端用户增强的隐私保护。然而，改进空间也很明显，特别是在影响用户主导的数据最小化结果的指标方面。

## 7.2 用户主导的数据最小化弥合 隐私意识与控制之间的鸿沟

我们的研究探讨了人们对使用Chat-GPT的隐私顾虑，重点关注信息可识别性与缺乏透明度问题。参与者在日常使用中采取了一些数据最小化策略，但认为这些措施不足，导致他们感到脆弱且缺乏控制感（见6.1小节）。这些发现印证了为用户提供便捷灵活的数据最小化支持的必要性，同时也更深入地揭示了Rescriber的某些设计选择如何有效应对用户个性化的风险认知。例如，赋予用户完全掌控清理操作的权限受到用户好评（见6.2.6小节），还能帮助他们根据个人偏好管理隐私-效用权衡（如同案例中对可识别性的差异化评估，参见6.1.1和6.4.1小节）。用户尤其看重我们工具提供的流程优化功能，包括替代解决方案（6.3.4小节）、“全选”和“写回”等便利功能（6.3.2小节），以及通过针对不同场景提供抽象化与替换两种清理选项，帮助用户在隐私与效用间找到平衡点（6.4.2小节）。

我们的工具解决的另一个缺口是提升用户对敏感信息披露的认知（6.5.1小节）。参与者频繁提到该工具的教育价值，认为其与增强的隐私控制和保护同等重要。尽管并非专门设计为教育工具，但我们的工具通过提供现场学习机会，展示了实现隐私教育的一个示例。

常规数据共享中的隐私风险，类似于体验式学习[16]。用户愿意参与探索不同的清理选项，因为它能提供即时且切实的益处。这一洞察力可为更多基于用户主导数据最小化背景下的隐私教育设计与研究提供参考。

## 7.3 人工智能驱动设计的启示 人在回路隐私控制

我们的方法强调，需要从传统的人类不在回路的隐私增强技术（PETs）转向支持人在回路、细粒度隐私管理的工具。

虽然人工智能有助于实现这一目标，但将其妥善整合到此类系统中，使用户能够信任它并愿意使用它，并且发现它提供的保护符合他们的期望，需要深入理解人类与人工智能之间互动的动态机制，这些互动由它们之间的界面所调节。下文我们将反思从设计此类系统和界面的研究中汲取的教训。

迈向以人为中心的评估。尽管准确性、精确率和召回率等性能指标都是评估模型性能的重要方式，但它们对用户体验和决策的影响却大不相同。如第6.2小节所述，检测与清理结果的全面性和一致性是影响用户信任的关键因素。

具体而言，用户能够容忍检测结果中的误报（例如较不敏感的敏感信息）和错误标记的数据类别，而单个漏报则可能对其系统信任度造成灾难性影响。这表明某些指标（如召回率）比其他指标（如精确率）更为重要。

此外，我们发现用户对敏感信息泄露的认知涵盖广泛主题，这些主题超出了我们分类法的范围，且可能无法被任何分类法完全涵盖（见第6.2.4小节）。当Rescriber-Llama3-8B模型生成超出分类法的检测结果（如体重和身高）时，用户实际上更青睐这些结果，而非严格遵循分类法、未将其识别为敏感数据的Rescriber-GPT-4o。然而，基于分类法的常见基准测试方法会在结果中曲解用户对此类案例的偏好，这揭示了基于分类法的评估方法的局限性，并强调了超越个人信息（PII）[6]的隐私保护的重要性。

全面性与控制之间的平衡。在6.2.6小节和6.2.5小节中，我们观察到关于隐私控制看似矛盾的需求。在6.3.1小节中，我们还发现用户由于对模型能力的深入见解不足，难以在隐私与效用之间找到最佳权衡。这种张力凸显了在赋予用户建立主观信任的控制权与确保全面数据最小化之间取得平衡的必要性，同时避免因用户有限的经验或注意力跨度而受到削弱。未来研究可以进一步探索不同层次的用户参与度和能动性，以及个体差异的影响，以设计出满足这两项重要要求的更好界面。

**Consistency, predictability, and learning curve.** Users reported abstraction being useful for retaining contexts while conveying general information. However, it was used less often than expected, partly due to the lack of predictability in abstraction results. This is exacerbated by the inconsistent results between partial and complete input (both versions) and across different trials (**Rescriber-GPT-4o** only). As a result, the learning curve is heightened. Additionally, predictability also involves aligning users' expected and actual abstraction granularity; overgeneralization is a main reason for dissatisfaction with abstraction results.

**Transparency and trust.** Despite the positive feedback, users also expressed concerns about “*offsetting the privacy concern away from ChatGPT onto this extension*” (P4). They stressed the importance of high transparency, expecting the tool to run locally. This highlights that users may expect higher trust standards from privacy-protection tools.

## 8 Conclusion

In this work, we developed a system powered by LLMs that allows users to sanitize prompts, enabling user-led data minimization for LLM-based conversational agents. Our study revealed that most privacy concerns stem from the fear of being identifiable through their messages and the potential risks that follow. By offering quick, easy-to-use sanitization options, users could achieve satisfactory responses from GPT while feeling secure about their privacy. Notably, users became more aware of privacy considerations and actively evaluated whether exposing certain information was necessary for an appropriate response. Our findings also show no significant difference in user perception between the cloud-based and on-device models, suggesting that smaller LLMs offer a promising solution for on-device, user-facing privacy management in LLM-based applications.

## Acknowledgments

The project is in part supported by a gift from Google on “User-Centered Privacy Design in Smart Mobile Text Entry”. We are thankful to Weiyan Shi, Jackie Yang, Niloofar Miresghallah, and all members of the PEACH lab for their helpful suggestions and feedback at different stages of this project.

## References

- [1] 2024. Private Cloud Compute: A new frontier for AI privacy in the cloud. <https://security.apple.com/blog/private-cloud-compute/> Accessed: 2024-09-07.
- [2] AI4Privacy. 2023. PII Masking 200K Dataset. <https://huggingface.co/datasets/ai4privacy/pii-masking-200k> Accessed: 2024-09-07.
- [3] Hugh Beyer and Karen Holtzblatt. 1999. Contextual design. *interactions* 6, 1 (1999), 32–42.
- [4] Asia J. Biega, Peter Potash, Hal Daumé, Fernando Diaz, and Michèle Finck. 2020. Operationalizing the Legal Principle of Data Minimization for Personalization. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. ACM, 399–408. doi:10.1145/3397271.3401034
- [5] Hee-Seung Henry Bom. 2023. Exploring the Opportunities and Challenges of ChatGPT in Academic Writing: a Roundtable Discussion. *Nuclear Medicine and Molecular Imaging* 57, 4 (May 2023), 165–167. doi:10.1007/s13139-023-00809-2
- [6] Hannah Brown, Katherine Lee, Fatemehsadat Miresghallah, Reza Shokri, and Florian Tramér. 2022. What Does It Mean for a Language Model to Preserve Privacy? In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. ACM, 2280–2292. doi:10.1145/3531146.3534642
- [7] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).
- [8] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646* (2022).
- [9] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting Training Data from Large Language Models.. In *USENIX Security Symposium*, Vol. 6.
- [10] Szu-Wei Cheng, Chung-Wen Chang, Wan-Jung Chang, Hao-Wei Wang, Chih-Sung Liang, Taishiro Kishimoto, Jane Pei-Chen Chang, John S Kuo, and Kuan-Pin Su. 2023. The now and future of ChatGPT and GPT in psychiatry. *Psychiatry and clinical neurosciences* 77, 11 (2023), 592–596.
- [11] Chun Jie Chong, Chenxi Hou, Zhihao Yao, and Seyed Mohammadjavad Seyed Talebi. 2024. Casper: Prompt Sanitization for Protecting User Privacy in Web-Based Large Language Models. *arXiv preprint arXiv:2408.07004* (2024).
- [12] Youngwook Do, Nivedita Arora, Ali Mirzazadeh, Injoo Moon, Eryue Xu, Zhihan Zhang, Gregory D Abowd, and Sauvik Das. 2023. Powering for privacy: improving user trust in smart speaker microphones with intentional powering and perceptible assurance. In *32nd USENIX Security Symposium (USENIX Security 23)*, 2473–2490.
- [13] Youngwook Do, Jung Wook Park, Yuxi Wu, Avinandan Basu, Dingtian Zhang, Gregory D. Abowd, and Sauvik Das. 2021. Smart Webcam Cover: Exploring the Design of an Intelligent Webcam Cover to Improve Usability and Trust: Exploring the Design of an Intelligent Webcam Cover to Improve Usability and Trust. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (December 2021), 1–21. doi:10.1145/3494983
- [14] Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, and Wei Xu. 2024. Reducing Privacy Risks in Online Self-Disclosures with Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1372–13754. doi:10.18653/v1/2024.acl-long.741
- [15] Federal Privacy Council. 2023. Fair Information Practice Principles (FIPPs). <https://www.fpc.gov/resources/fipps/> Accessed: 2024-09-07.
- [16] James W Gentry. 1990. What is experiential learning. *Guide to business gaming and experiential learning* 9, 1 (1990), 20–32.
- [17] Google. 2023. Introducing Gemini: our largest and most capable AI model. <https://blog.google/technology/ai/google-gemini-ai/> Accessed: 2024-09-07.
- [18] Joel Jang, Dongkeun Yoon, Soheer 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).
- [19] Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*. PMLR, 10697–10707.
- [20] Nir Kshetri. 2023. Cybercrime and Privacy Threats of Large Language Models. *IT Professional* 25, 3 (May 2023), 9–13. doi:10.1109/mitp.2023.3275489
- [21] James R. Lewis, Brian S. Utesch, and Deborah E. Maher. 2013. UMUX-LITE: when there's no time for the SUS: when there's no time for the SUS. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM. doi:10.1145/2470654.2481287
- [22] Tianshi Li, Yuvraj Agarwal, and Jason I. Hong. 2018. Coconut: An IDE Plugin for Developing Privacy-Friendly Apps: An IDE Plugin for Developing Privacy-Friendly Apps. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (December 2018), 1–35. doi:10.1145/3287056
- [23] Tianshi Li, Sauvik Das, Hao-Ping (Hank) Lee, Dakuo Wang, Bingsheng Yao, and Zhiping Zhang. 2024. Human-Centred Privacy Research in the Age of Large Language Models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, 1–4. doi:10.1145/3613905.3643983
- [24] 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).
- [25] Yuanchun Li, Fanglin Chen, Toby Jia-Jun Li, Yao Guo, Gang Huang, Matthew Fredrikson, Yuvraj Agarwal, and Jason I. Hong. 2017. PrivacyStreams: Enabling Transparency in Personal Data Processing for Mobile Apps: Enabling Transparency in Personal Data Processing for Mobile Apps. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (September 2017), 1–26. doi:10.1145/3130941
- [26] Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelied. 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)*. Association for Computational Linguistics. doi:10.18653/v1/2021.acl-long.323
- [27] Jimi Majumdar, Christophe Dupuy, Charith Peris, Sami Smaili, Rahul Gupta, and Richard Zemel. 2022. Differentially private decoding in large language models. *arXiv preprint arXiv:2205.13621* (2022).
- [28] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023.

一致性、可预测性与学习曲线。用户反馈抽象化有助于在传达通用信息的同时保留上下文。然而，其使用频率低于预期，部分原因是抽象结果缺乏可预测性。这一问题因部分和完整输入（两个版本）之间以及不同试验（仅限Rescriber-GPT-4o）结果的不一致性而加剧。因此，学习曲线被抬高。此外，可预测性还涉及与用户预期的对齐。

预期与实际抽象化粒度；过度泛化是导致对抽象结果不满的主要原因。

透明度与信任。尽管反馈积极，用户仍对“将隐私担忧从ChatGPT转移至此扩展”（P4）表示忧虑。他们强调高透明度的重要性，期望该工具能在本地运行。

这突显出用户可能对隐私保护工具抱有更高的信任标准。

## 8 结论

在这项工作中，我们开发了一个由大型语言模型驱动的系统，允许用户清理提示，实现用户主导的数据最小化，用于基于LLM的对话代理。我们的研究表明，大多数隐私顾虑源于担心通过消息被识别以及随之而来的潜在风险。通过提供快速、易于使用的清理选项，用户可以从GPT获得满意的响应，同时对其隐私感到安全。值得注意的是，

用户对隐私考量的意识增强，并积极评估暴露某些信息是否对获得适当响应是必要的。我们的研究结果还显示，在用户感知上，云端和本地模型之间没有显著差异，这表明小型大语言模型为基于LLM的应用提供了设备端、面向用户的隐私管理的有前景的解决方案。

## 致谢

该项目部分由谷歌关于“智能移动文本输入中的用户中心隐私设计”的赠款支持。我们感谢施伟岩、杨杰基、尼卢法尔·米雷斯加拉以及PEACH实验室所有成员在项目不同阶段提供的宝贵建议和反馈。

## 参考文献

- [1] 2024. Private Cloud Compute: A new frontier for AI privacy in the cloud. <https://security.apple.com/blog/private-cloud-compute/> Accessed: 2024-09-07.
- [2] AI4Privacy. 2023. PII Masking 200K Dataset. <https://huggingface.co/datasets/ai4privacy/pii-masking-200k> Accessed: 2024-09-07.
- [3] Hugh Beyer and Karen Holtzblatt. 1999. Contextual design. *interactions* 6, 1 (1999), 32–42.
- [4] Asia J. Biega, Peter Potash, Hal Daumé, Fernando Diaz, and Michèle Finck. 2020. Operationalizing the Legal Principle of Data Minimization for Personalization. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. ACM, 399–408. doi:10.1145/3397271.3401034
- [5] Hee-Seung Henry Bom. 2023. Exploring the Opportunities and Challenges of ChatGPT in Academic Writing: a Roundtable Discussion. *Nuclear Medicine and Molecular Imaging* 57, 4 (May 2023), 165–167. doi:10.1007/s13139-023-00809-2
- [6] Hannah Brown, Katherine Lee, Fatemehsadat Miresghallah, Reza Shokri, and Florian Tramér. 2022. What Does It Mean for a Language Model to Preserve Privacy? In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. ACM, 2280–2292. doi:10.1145/3531146.3534642
- [7] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).
- [8] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646* (2022).
- [9] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting Training Data from Large Language Models.. In *USENIX Security Symposium*, Vol. 6.
- [10] Szu-Wei Cheng, Chung-Wen Chang, Wan-Jung Chang, Hao-Wei Wang, Chih-Sung Liang, Taishiro Kishimoto, Jane Pei-Chen Chang, John S Kuo, and Kuan-Pin Su. 2023. The now and future of ChatGPT and GPT in psychiatry. *Psychiatry and clinical neurosciences* 77, 11 (2023), 592–596.
- [11] Chun Jie Chong, Chenxi Hou, Zhihao Yao, and Seyed Mohammadjavad Seyed Talebi. 2024. Casper: Prompt Sanitization for Protecting User Privacy in Web-Based Large Language Models. *arXiv preprint arXiv:2408.07004* (2024).
- [12] Youngwook Do, Nivedita Arora, Ali Mirzazadeh, Injoo Moon, Eryue Xu, Zhihan Zhang, Gregory D Abowd, and Sauvik Das. 2023. Powering for privacy: improving user trust in smart speaker microphones with intentional powering and perceptible assurance. In *32nd USENIX Security Symposium (USENIX Security 23)*, 2473–2490.
- [13] Youngwook Do, Jung Wook Park, Yuxi Wu, Avinandan Basu, Dingtian Zhang, Gregory D. Abowd, and Sauvik Das. 2021. Smart Webcam Cover: Exploring the Design of an Intelligent Webcam Cover to Improve Usability and Trust: Exploring the Design of an Intelligent Webcam Cover to Improve Usability and Trust. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (December 2021), 1–21. doi:10.1145/3494983
- [14] Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, and Wei Xu. 2024. Reducing Privacy Risks in Online Self-Disclosures with Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1372–13754. doi:10.18653/v1/2024.acl-long.741
- [15] Federal Privacy Council. 2023. Fair Information Practice Principles (FIPPs). <https://www.fpc.gov/resources/fipps/> Accessed: 2024-09-07.
- [16] James W Gentry. 1990. What is experiential learning. *Guide to business gaming and experiential learning* 9, 1 (1990), 20–32.
- [17] Google. 2023. Introducing Gemini: our largest and most capable AI model. <https://blog.google/technology/ai/google-gemini-ai/> Accessed: 2024-09-07.
- [18] Joel Jang, Dongkeun Yoon, Soheer 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).
- [19] Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*. PMLR, 10697–10707.
- [20] Nir Kshetri. 2023. Cybercrime and Privacy Threats of Large Language Models. *IT Professional* 25, 3 (May 2023), 9–13. doi:10.1109/mitp.2023.3275489
- [21] James R. Lewis, Brian S. Utesch, and Deborah E. Maher. 2013. UMUX-LITE: when there's no time for the SUS: when there's no time for the SUS. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM. doi:10.1145/2470654.2481287
- [22] Tianshi Li, Yuvraj Agarwal, and Jason I. Hong. 2018. Coconut: An IDE Plugin for Developing Privacy-Friendly Apps: An IDE Plugin for Developing Privacy-Friendly Apps. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (December 2018), 1–35. doi:10.1145/3287056
- [23] Tianshi Li, Sauvik Das, Hao-Ping (Hank) Lee, Dakuo Wang, Bingsheng Yao, and Zhiping Zhang. 2024. Human-Centred Privacy Research in the Age of Large Language Models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, 1–4. doi:10.1145/3613905.3643983
- [24] 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).
- [25] Yuanchun Li, Fanglin Chen, Toby Jia-Jun Li, Yao Guo, Gang Huang, Matthew Fredrikson, Yuvraj Agarwal, and Jason I. Hong. 2017. PrivacyStreams: Enabling Transparency in Personal Data Processing for Mobile Apps: Enabling Transparency in Personal Data Processing for Mobile Apps. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (September 2017), 1–26. doi:10.1145/3130941
- [26] Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelied. 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)*. Association for Computational Linguistics. doi:10.18653/v1/2021.acl-long.323
- [27] Jimi Majumdar, Christophe Dupuy, Charith Peris, Sami Smaili, Rahul Gupta, and Richard Zemel. 2022. Differentially private decoding in large language models. *arXiv preprint arXiv:2205.13621* (2022).
- [28] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023.

- Model. *Information Systems Research* 15, 4 (December 2004), 336–355. doi:10.1287/isre.1040.0032
- [29] Ethan Mendes, Yang Chen, James Hays, Sauvik Das, Wei Xu, and Alan Ritter. 2024. Granular Privacy Control for Geolocation with Vision Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 17240–17292. doi:10.18653/v1/2024.emnlp-main.957
- [30] Niloofar Mireshghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. 2024. Trust No Bot: Discovering Personal Disclosures in Human-LLM Conversations in the Wild. arXiv:2407.11438 [cs.CL] <https://arxiv.org/abs/2407.11438>
- [31] Niloofar Mireshghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. 2024. Trust No Bot: Discovering Personal Disclosures in Human-LLM Conversations in the Wild. arXiv preprint arXiv:2407.11438 (2024).
- [32] Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. 2023. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. arXiv preprint arXiv:2310.17884 (2023).
- [33] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramér, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models. arXiv preprint arXiv:2311.17035 (2023).
- [34] Helen Nissenbaum. 2004. Privacy as contextual integrity. *Wash. L. Rev.* 79 (2004), 119.
- [35] Annika Willoch Olstad, Anthi Papadopoulou, and Pierre Lison. 2023. Generation of Replacement Options in Text Sanitization. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*. 292–300.
- [36] OpenAI. 2022. Introducing ChatGPT. <https://openai.com/index/chatgpt/> Accessed: 2024-09-07.
- [37] European Parliament and Council of the European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). <https://eur-lex.europa.eu/eli/reg/2016/679/oj> Accessed: 2024-09-07.
- [38] Charith Peris, Christophe Dupuy, Jimit Majmudar, Rahil Parikh, Sami Smaili, Richard Zemel, and Rahul Gupta. 2023. Privacy in the Time of Language Models. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM '23)*. ACM, 1291–1292. doi:10.1145/3539597.3575792
- [39] Bashir Rastegarpanah, Krishna Gummadi, and Mark Crovella. 2021. Auditing black-box prediction models for data minimization compliance. *Advances in Neural Information Processing Systems* 34 (2021), 20621–20632.
- [40] RyokoAI. 2023. ShareGPT52K Dataset. <https://huggingface.co/datasets/RyokoAI/ShareGPT52K> Accessed: 2024-09-07.
- [41] Peter Schaar. 2010. Privacy by design. *Identity in the Information Society* 3, 2 (2010), 267–274.
- [42] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2024. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems* 36 (2024).
- [43] Awanthika Senarath and Nalin Asanka Gamagedara Arachchilage. 2018. Understanding software developers' approach towards implementing data minimization. arXiv preprint arXiv:1808.01479 (2018).
- [44] Yijia Shao, Tianshi Li, Weiyan Shi, Yanchen Liu, and Diyi Yang. 2024. PrivacyLens: Evaluating Privacy Norm Awareness of Language Models in Action. arXiv preprint arXiv:2409.00138 (2024).
- [45] Tanusree Sharma, Lin Kyi, Yang Wang, and Asia J Biega. 2024. "I'm not convinced that they don't collect more than is necessary":{User-Controlled} Data Minimization Design in Search Engines. In *33rd USENIX Security Symposium (USENIX Security 24)*. 2797–2812.
- [46] Adam Shostack. 2024. The Boy Who Survived: Removing Harry Potter from an LLM is harder than reported. arXiv:2403.12082 [cs.CL]
- [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] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023).
- [49] Florian Tramer and Dan Boneh. 2018. Slalom: Fast, verifiable and private execution of neural networks in trusted hardware. arXiv preprint arXiv:1806.03287 (2018).
- [50] Dimusha Vatsalan, Peter Christen, and Vassilios S Verykios. 2013. A taxonomy of privacy-preserving record linkage techniques. *Information Systems* 38, 6 (2013), 946–969.
- [51] Ari Ezra Waldman. 2018. *Privacy as trust: Information privacy for an information age*. Cambridge University Press.
- [52] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. arXiv preprint arXiv:2206.07682 (2022).
- [53] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. 2021. Differentially private fine-tuning of language models. arXiv preprint arXiv:2110.06500 (2021).
- [54] Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramér, and Nicholas Carlini. 2021. Counterfactual memorization in neural language models. arXiv preprint arXiv:2112.12938 (2021).
- [55] Lotus Zhang, Abigale Stangl, Tanusree Sharma, Yu-Yun Tseng, Inan Xu, Danna Gurari, Yang Wang, and Leah Findlater. 2024. Designing Accessible Obfuscation Support for Blind Individuals' Visual Privacy Management. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, 1–19. doi:10.1145/3613904.3642713
- [56] Zhiping Zhang, Bingcan Guo, and Tianshi Li. 2024. Privacy Leakage Overshadowed by Views of AI: A Study on Human Oversight of Privacy in Language Model Agent. arXiv preprint arXiv:2411.01344 (2024).
- [57] Zhiping Zhang, Michelle Jia, Hao-Ping (Hank) 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 (CHI '24)*. ACM, 1–26. doi:10.1145/3613904.3642385
- [58] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuany Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2024. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (New Orleans, LA, USA) (NIPS '23)*. Curran Associates Inc, Red Hook, NY, USA, Article 2020, 29 pages.

## A System Benchmarking and On-Device Feasibility Study

### A.1 Time Performance Benchmark

We present the full results of our experiments benchmarking the time performance of Rescriber-GPT-4o, Rescriber-Llama3-8B, and Presidio. For the on-device models, we ran the experiments with four varying device configurations, including three consumer computers: Apple M2 (24 GB memory), Apple M4 (16GB memory), Apple M1 Max (64GB memory), as well as one cloud virtual machine with an NVIDIA L4 GPU (24GB VRAM).

**Table 8: Comparison of Response Times Among Presidio, Rescriber-GPT-4o and Rescriber-Llama3-8B. On-device models are benchmarked with different device configurations.**

Model	Device	Min (s)	Max (s)	Mean (s)	Std (s)
Rescriber-GPT-4o		0.53	3.98	1.41	0.65
Rescriber-Llama3-8B (first detection)	M2, 24GB	1.34	4.52	2.01	0.48
Rescriber-Llama3-8B (full detection)	M2, 24GB	1.62	7.93	3.40	1.29
Rescriber-Llama3-8B (first detection)	M4, 16GB	1.07	3.58	1.57	0.38
Rescriber-Llama3-8B (full detection)	M4, 16GB	1.23	6.35	2.66	1.02
Rescriber-Llama3-8B (first detection)	M1 Max, 64GB	0.45	1.60	0.66	0.16
Rescriber-Llama3-8B (full detection)	M1 Max, 64GB	0.54	2.58	1.09	0.41
Rescriber-Llama3-8B (first detection)	L4, 24GB	0.46	1.22	0.67	0.15
Rescriber-Llama3-8B (full detection)	L4, 24GB	0.52	2.87	1.16	0.47
Presidio	M2, 24GB	0.6	1.98	0.73	0.19
Presidio	M4, 16GB	0.48	0.82	0.53	0.06
Presidio	M1 Max, 64GB	0.74	9.64	0.84	0.53
Presidio	L4, 24GB	1.56	8.43	1.73	0.47

### A.2 On-Device Feasibility Study Information

We collected user demographic and device information to ensure participants met the study requirements and represented a diverse range of backgrounds and device configurations. See Table 9 for more detail.

- Model. *Information Systems Research* 15, 4 (December 2004), 336–355. doi:10.1287/isre.1040.0032
- [29] Ethan Mendes, Yang Chen, James Hays, Sauvik Das, Wei Xu, and Alan Ritter. 2024. Granular Privacy Control for Geolocation with Vision Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 17240–17292. doi:10.18653/v1/2024.emnlp-main.957
- [30] Niloofar Mireshghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. 2024. Trust No Bot: Discovering Personal Disclosures in Human-LLM Conversations in the Wild. arXiv:2407.11438 [cs.CL] <https://arxiv.org/abs/2407.11438>
- [31] Niloofar Mireshghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. 2024. Trust No Bot: Discovering Personal Disclosures in Human-LLM Conversations in the Wild. arXiv preprint arXiv:2407.11438 (2024).
- [32] Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. 2023. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. arXiv preprint arXiv:2310.17884 (2023).
- [33] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramér, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models. arXiv preprint arXiv:2311.17035 (2023).
- [34] Helen Nissenbaum. 2004. Privacy as contextual integrity. *Wash. L. Rev.* 79 (2004), 119.
- [35] Annika Willoch Olstad, Anthi Papadopoulou, and Pierre Lison. 2023. Generation of Replacement Options in Text Sanitization. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*. 292–300.
- [36] OpenAI. 2022. Introducing ChatGPT. <https://openai.com/index/chatgpt/> Accessed: 2024-09-07.
- [37] European Parliament and Council of the European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). <https://eur-lex.europa.eu/eli/reg/2016/679/oj> Accessed: 2024-09-07.
- [38] Charith Peris, Christophe Dupuy, Jimit Majmudar, Rahil Parikh, Sami Smaili, Richard Zemel, and Rahul Gupta. 2023. Privacy in the Time of Language Models. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM '23)*. ACM, 1291–1292. doi:10.1145/3539597.3575792
- [39] Bashir Rastegarpanah, Krishna Gummadi, and Mark Crovella. 2021. Auditing black-box prediction models for data minimization compliance. *Advances in Neural Information Processing Systems* 34 (2021), 20621–20632.
- [40] RyokoAI. 2023. ShareGPT52K Dataset. <https://huggingface.co/datasets/RyokoAI/ShareGPT52K> Accessed: 2024-09-07.
- [41] Peter Schaar. 2010. Privacy by design. *Identity in the Information Society* 3, 2 (2010), 267–274.
- [42] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2024. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems* 36 (2024).
- [43] Awanthika Senarath and Nalin Asanka Gamagedara Arachchilage. 2018. Understanding software developers' approach towards implementing data minimization. arXiv preprint arXiv:1808.01479 (2018).
- [44] Yijia Shao, Tianshi Li, Weiyan Shi, Yanchen Liu, and Diyi Yang. 2024. PrivacyLens: Evaluating Privacy Norm Awareness of Language Models in Action. arXiv preprint arXiv:2409.00138 (2024).
- [45] Tanusree Sharma, Lin Kyi, Yang Wang, and Asia J Biega. 2024. "I'm not convinced that they don't collect more than is necessary":{User-Controlled} Data Minimization Design in Search Engines. In *33rd USENIX Security Symposium (USENIX Security 24)*. 2797–2812.
- [46] Adam Shostack. 2024. The Boy Who Survived: Removing Harry Potter from an LLM is harder than reported. arXiv:2403.12082 [cs.CL]
- [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] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023).
- [49] Florian Tramer and Dan Boneh. 2018. Slalom: Fast, verifiable and private execution of neural networks in trusted hardware. arXiv preprint arXiv:1806.03287 (2018).
- [50] Dimusha Vatsalan, Peter Christen, and Vassilios S Verykios. 2013. A taxonomy of privacy-preserving record linkage techniques. *Information Systems* 38, 6 (2013), 946–969.
- [51] Ari Ezra Waldman. 2018. *Privacy as trust: Information privacy for an information age*. Cambridge University Press.
- [52] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. arXiv preprint arXiv:2206.07682 (2022).
- [53] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. 2021. Differentially private fine-tuning of language models. arXiv preprint arXiv:2110.06500 (2021).
- [54] Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramér, and Nicholas Carlini. 2021. Counterfactual memorization in neural language models. arXiv preprint arXiv:2112.12938 (2021).
- [55] Lotus Zhang, Abigale Stangl, Tanusree Sharma, Yu-Yun Tseng, Inan Xu, Danna Gurari, Yang Wang, and Leah Findlater. 2024. Designing Accessible Obfuscation Support for Blind Individuals' Visual Privacy Management. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, 1–19. doi:10.1145/3613904.3642713
- [56] Niloofar Mireshghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. 2024. Trust No Bot: Discovering Personal Disclosures in Human-LLM Conversations in the Wild. arXiv:2407.11438 [cs.CL] <https://arxiv.org/abs/2407.11438>
- [57] Zhiping Zhang, Bingcan Guo, and Tianshi Li. 2024. Privacy Leakage Overshadowed by Views of AI: A Study on Human Oversight of Privacy in Language Model Agent. arXiv preprint arXiv:2411.01344 (2024).
- [58] Zhiping Zhang, Michelle Jia, Hao-Ping (Hank) 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 (CHI '24)*. ACM, 1–26. doi:10.1145/3613904.3642385
- [59] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuany Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2024. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (New Orleans, LA, USA) (NIPS '23)*. Curran Associates Inc, Red Hook, NY, USA, Article 2020, 29 pages.

2022. Emergent abilities of large language models. arXiv preprint arXiv:2206.07682 (2022).

[53] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. 2021. Differentially private fine-tuning of language models. arXiv preprint arXiv:2110.065

**Table 9: Participant and Device information of the On-Device Feasibility Testing Study**

ID	Gender	Age Group	Background	Chip	Memory (GB)
F1	Male	25-34	Tech	M1 Max	64
F2	Male	25-34	Tech	M1 Pro	16
F3	Female	18-24	Non-tech	M1 Pro	16
F4	Female	18-24	Non-tech	M2	16
F5	Female	18-24	Non-tech	M4	16
F6	Female	18-24	Non-tech	M1	16
F7	Female	25-34	Mixed	M1	16
F8	Female	18-24	Mixed	M3 Pro	36

## B Interview Protocol

### B.1 Pre-Interview Message

Thank you for your willingness to join our study! During the study, we will test our privacy-preserving extension together. Specifically, the researcher will ask you to provide real-time feedback while you use our extension during a conversation with ChatGPT.

(Make sure the participant received the email with instructions and survey link.)

Here are a few things to set you ready!

- (1) Please fill out the survey so we can better prepare for your experience. It should take less than 3 minutes to finish.
- (2) Please prepare a ChatGPT prompt to test beforehand. We highly suggest you review your ChatGPT chat history and pick a past conversation that:
  - Contains personal information that you are concerned about sharing with ChatGPT.
  - You are comfortable showcasing for the study purpose.
- (3) Feel free to take a look at the participant information sheet for more information about the study. All your personal information will be kept confidential.

(If the participant has questions about how to select a ChatGPT conversation, provide guidance based on their concerns about privacy.)

### B.2 Counterbalancing Settings

The study involves 12 participants across 4 conditions, where each participant interacts with two models (Model 1 and Model 2), using both their own and provided examples. The order of testing is randomized to avoid bias. See Table 10.

### B.3 Interview Script

Hi there! We are researchers from Northeastern University, and we are interested in learning about your privacy concerns when using ChatGPT. During this interview, we will test a privacy-preserving extension together. Specifically, I will ask you to provide feedback while you use our extension during conversations with ChatGPT.

(Send the information sheet to the participant before starting the conversation.)

Here is the study information sheet. During the study, you are free to answer only the questions you're comfortable with. If there's a question that you don't want to discuss, just let me know. It won't affect your compensation.

**Table 10: Counterbalancing settings**

ID	Example	First Example	Model Order
P1	E1	Participant's Example	Model 1 → Model 2
P2	E1	Participant's Example	Model 2 → Model 1
P3	E2	Participant's Example	Model 1 → Model 2
P4	E2	Participant's Example	Model 2 → Model 1
P5	E3	Participant's Example	Model 1 → Model 2
P6	E3	Participant's Example	Model 2 → Model 1
P7	E1	Experimenter's Example	Model 1 → Model 2
P8	E1	Experimenter's Example	Model 2 → Model 1
P9	E2	Experimenter's Example	Model 1 → Model 2
P10	E2	Experimenter's Example	Model 2 → Model 1
P11	E3	Experimenter's Example	Model 1 → Model 2
P12	E3	Experimenter's Example	Model 2 → Model 1

**表9：设备端可行性测试研究的参与者与设备信息  
可行性测试研究**

ID	性别	年龄段	背景	Chip	内存 (GB)
F1	Male	25-34	Tech	M1 Max	64
F2	Male	25-34	Tech	M1 Pro	16
F3	女性	18-24	非技术	M1 Pro	16
F4	女性	18-24	非技术	M2	16
F5	女性	18-24	非技术	M4	16
F6	女性	18-24	非技术	M1	16
F7	女性	25-34	混合	M1	16
F8	女性	18-24	混合	M3 Pro	36

## B访谈协议

### B.1 访谈前信息

感谢您愿意加入我们的研究！在研究期间，我们将共同测试我们的隐私保护扩展。具体来说，研究会要求您在与ChatGPT对话时使用我们的扩展程序，并提供实时反馈。

(确保参与者已收到包含说明的电子邮件及调查问卷链接。)

以下是一项准备工作！

- (1) 请填写调查问卷，以便我们更好地为您体验做准备。完成问卷应不超过3分钟。
- (2) 请提前准备一个ChatGPT提示用于测试。我们强烈建议您审查您的ChatGPT聊天记录并选择一段过往对话，该对话需满足以下条件：
  - 包含您担忧与ChatGPT共享的个人信息。
  - 关于与ChatGPT共享。
  - 您愿意为研究目的展示该对话内容。
- (3) 可随时查阅参与者信息表，如需了解更多关于该研究的信息。您所有的个人信息都将被保密。

(如果参与者对如何选择Chat-GPT对话有疑问，请根据他们对隐私的担忧提供指导。)

### B.2 平衡设置

该研究涉及12名参与者，分为4种条件，每位参与者需与两个模型（模型1和模型2）进行交互，使用包括他们自己提供的示例和给定的示例。测试顺序是随机化的以避免偏见。参见表10。

### B.3 访谈脚本

你好！我们是来自东北大学的研究人员，我们有兴趣了解您在使用ChatGPT时的隐私顾虑。在这次访谈中，我们将测试一个隐私保护的共同扩展。具体来说，我会请您在使用我们的扩展时提供反馈当您在与ChatGPT对话过程中使用我们的扩展时。

(在开始对话前，请将研究信息表发送给参与者。)

这是研究信息表。在研究过程中，您可以自由回答您感到舒适的问题。如果有遇到您不想讨论的问题，请直接告知。这不会影响您的补偿。

**表10：平衡设置**

ID	示例	第一个示例	模型顺序
P1	E1	参与者示例	模型1 → 模型2
P2	E1	参与者示例	模型2 → 模型1
P3	E2	参与者示例	模型1 → 模型2
P4	E2	参与者示例	模型2 → 模型1
P5	E3	参与者示例	模型1 → 模型2
P6	E3	参与者示例	模型2 → 模型1
P7	E1	实验者示例	模型1 → 模型2
P8	E1	实验者示例	模型2 → 模型1
P9	E2	实验者示例	模型1 → 模型2
P10	E2	实验者示例	模型2 → 模型1
P11	E3	实验者示例	模型1 → 模型2
P12	E3	实验者示例	模型2 → 模型1

(确保参与者理解补偿方法：  
亚马逊礼品卡或Prolific支付。)

我还将视频录制本次会议用于笔记记录目的。您的所有响应都将保密。在研究过程中，您可以随时关闭摄像头。这样可以吗？

(如果对方回答是，则开始录制会话。)  
太好了！我们开始吧。

### B.4 一般问题

首先，我想了解您通常使用ChatGPT做什么。  
能否举一个最近的典型示例？

(注意参与者的响应，并准备好进一步探讨  
基于他们的使用习惯。)

[如果他们曾尝试隐藏PII] 我在调查问卷中看到您选择了“是”关于尽量减少与ChatGPT分享不必要的信息，您能详细说明在哪些情况下采取了哪些措施吗？

[如果他们未曾尝试隐藏PII] 我在调查问卷中注意到您对与ChatGPT分享个人信息有所顾虑，  
能否详细说明这些情况？

这是个很好的示例！太棒了，我们的团队已经开发了一个工具，  
这是一款Chrome扩展程序，有助于识别  
您的个人信息，并用通用术语进行脱敏处理，同时  
使用GPT。我们诚挚邀请您试用并提供  
反馈，请按照以下我的指示操作。

我们希望您能通过一次真实对话来测试该扩展——这段对话  
将涉及您与ChatGPT分享的个人身份信息、个人经历或其  
他敏感信息。若您在研究前已准备好此类对话，可直接用于测  
试；若未准备也无需担心，您可花几分钟回顾ChatGPT对话  
历史，选择一段愿意与我们分享的对话。

让用户将示例复制粘贴到空白文档页面，以便  
以便我们都能审查它。

太好了！能否向我们展示这段对话并简要说明  
您选择它的原因？

谢谢！现在我们将播放一段演示视频向您展示我们的扩展  
工作原理：

[Play demo video](#)

## B.5 Contextual Experience User Testing

We designed two versions of the systems to test the efficacy of different models for the privacy information masking task. To avoid biasing you when testing them, we will share details of how the two models work at the end of the interview.

We would love you to test 2 conversations with 2 different models of the extension, so in total, 4 sessions with ChatGPT using our extension. One is the real-world conversation you prepared and just demonstrated to us; another is an example we prepared. Please interact with ChatGPT as you would do in a realistic setting, including asking follow-up questions to fulfill your request. We are all ears to hear about your feelings, including both positive and negative feedback. These will help us to iterate.

Session 1: Let's start with your personal example/our hypothetical example. You can find the conversation you prepared and use it to test the extension. Feel free to ask questions as we go.

(Guide participant through the process of using the extension for their first conversation.)

Session 2: Now, let's switch to the other model and try the same example again.

(Ensure the participant understands they need to test the same conversation with the other model.)

Session 3: Now, let's switch to our hypothetical example/your personal example. I will send the prompt over to you. Feel free to ask questions if anything is unclear.

Choose one example from E1/E2/E3 based on the counterbalancing settings table

Session 4: Finally, let's test the hypothetical example again using the other model.

## B.6 Feedback Questions For Each Session

Based on your experience using the extension, I have a few questions:

- How satisfied are you with the response after using the extension?
- To what extent do you believe this experience helps protect your privacy? Why?
- What do you like or dislike about this experience? Why?

(Note any interesting behaviors or hesitations and ask follow-up questions as necessary.)

## B.7 Post-Session Survey

Please fill out this survey based on your experience: [survey link removed]

Please choose [A/B] for setting 1 and [E1/E2/E3/E4] for setting 2.

- A: Cloud-based (A)
- B: On-device (B)
- E1: Experimenter's example 1
- E2: Experimenter's example 2
- E3: Experimenter's example 3
- E4: Participant's own example

## B.8 Model Comparison After Session 2 And Session 4

Thanks for filling out the surveys! Let's briefly go over your responses for both models:

- The extension reduces the disclosure of unnecessary information to ChatGPT.
- The extension reduces the disclosure of my personal information to ChatGPT.
- I have fewer privacy concerns when using this extension.
- I would use this extension if it were well-designed for usability.

[If they gave the same score for both models] I noticed you gave the same score for both models. Could you explain why?

[If they gave different scores] I noticed you gave different scores for the models. Could you explain the differences you observed?

## B.9 Debrief

We designed the two versions of the systems to test the efficacy of different models for the privacy information masking task. We didn't share the details of how the two models work to avoid biasing you when answering these questions. Since you have finished all the tasks, I would like to share more information about the two models we used today.

In Model 1, personal information detection is implemented by prompting GPT-4. In Model 2, personal information detection is implemented by prompting a smaller large language model hosted on our server. In the GPT-4 powered model, OpenAI may securely retain your data for up to 30 days, and it will not use your data for training their models. In the self-hosted model, we don't store any data that you share during the study. If there is certain information you would like us to delete, please feel free to contact us and we will make our best efforts to delete it.

Do you have any questions?

## B.10 Closing Remarks

That's a wrap for our interview today! Thank you so much for taking the time to share your insights and experiences with us - it's been incredibly valuable. We hope you find it useful. Please don't hesitate to reach out if you have any further questions or thoughts you'd like to share. Thanks again for your contribution to our study!

## C Qualitative Code Book

Table 11: Qualitative Code Names and Sample Quotes

Code Name	Definition	Quotes
Feature: Replace Back	Replace placeholder text with PII in both messages sent by users and responses generated by GPT.	<i>Because realistically, when you're putting in personal information about yourself and you need it contextualized for specific use, even if it's something as simple as a name for readability and things like that</i>

[播放演示视频](#)

## B.5 情境体验用户测试

我们设计了两个版本的系统来测试不同模型在隐私信息屏蔽任务中的效能。为避免测试时产生偏见，我们将在访谈结束时说明两种模型的工作原理细节。

我们希望您测试扩展程序两种不同模型下的2组对话，因此总共需要通过我们的扩展程序与ChatGPT进行4次会话。一组是您准备并已向我们演示的真实对话，另一组是我们准备的示例。

请像在真实场景中那样与ChatGPT进行交互，包括提出后续问题以满足您的需求。我们洗耳恭听您的感受，无论是正面还是负面反馈。这些都将帮助我们不断改进。

会话1：让我们从您的个人示例/我们的假设示例开始。您可以找到准备好的对话并利用它

来测试扩展功能。过程中请随时提问。

(引导参与者完成使用扩展进行首次对话的流程。)

会话2：现在，让我们切换到另一模型并尝试相同的示例再次。

(确保参与者理解他们需要用同一对话测试另一模型。)

会话3：现在，让我们切换到我们的假设示例/您的个人示例。我将把提示发送给您。如有任何不清楚的地方，请随时提问。

根据平衡对照表，从示例1/示例2/示例3中选择一个示例设置表

会话4：最后，让我们使用另一模型再次测试该假设示例另一模型。

## B.6 各会话反馈问题

根据您使用该扩展的体验，我有几个问题：

- 使用该扩展后，您对响应的满意度如何？
- 您认为这次体验在多大程度上帮助保护了您的隐私？为什么？
- 您喜欢或不喜欢这次体验的哪些方面？为什么？

(注意任何有趣的行为或犹豫，并根据需要提出后续问题。)

## B.7 会后调查问卷

请根据您的经验填写这份调查问卷：[调查问卷链接移除]

请为设置1选择[A/B]，为设置选择[示例1/示例2/示例3/E4]

2.

- A: 基于云端 (A)
- B: 设备端 (B)
- E1: 实验者示例1
- E2: 实验者示例2
- E3: 实验者示例3
- E4: 参与者自选示例

## B.8 第2次和第4次会话后的模型比较

感谢您填写调查问卷！让我们简要回顾一下您对两个模型的反馈：

- 该扩展减少了对ChatGPT不必要信息的披露。
- 该扩展减少了对ChatGPT我的个人信息的披露。
- 使用此扩展时，我的隐私顾虑较少。
- 如果这个扩展在可用性方面设计良好，我会使用它。

[如果他们给两个模型打了相同的分数] 我注意到你给两个模型打了相同的分数。能解释一下原因吗？

[如果他们给了不同的分数] 我注意到你给了不同的分数给这些模型。能解释一下你观察到的差异吗？

## B.9 任务汇报

我们设计了两个版本的系统来测试不同模型在隐私信息屏蔽任务中的效能。为避免影响您的判断，我们未透露两种模型的具体运作细节。当您回答这些问题时。既然您已完成所有任务，我想分享更多关于这两个系统的信息。关于这些任务，我想分享更多关于这两者的信息。我们今天使用的模型。

在模型1中，个人信息检测通过提示GPT-4实现。在模型2中，个人信息检测是通过提示一个托管在我们服务器上的较小大型语言模型实现的。在由GPT-4驱动的模型中，OpenAI可能会安全地保留您的数据长达30天，并且不会将您的数据用于训练他们的模型。在自托管模型中，我们不会存储任何您在研究期间分享的数据。如果有某些信息您希望我们删除，请随时联系我们，我们将尽最大努力予以删除。

您有任何问题吗？

## B.10 结束语

今天的访谈就到这里结束了！非常感谢您抽出时间与我们分享您的见解和经验——这些对我们来说极其宝贵。希望您觉得这次交流有所帮助。请

随时联系我们，如果您还有任何其他问题或想法想要分享。再次感谢您对我们研究的贡献！

## C 定性代码手册

表11：定性代码名称与示例引用

代码名称	定义	引用
功能：替换回	将占位文本替换为用户发送的消息及GPT生成的响应中的个人身份信息。	因为实际上，当你输入关于自己的个人信息并需要将其置于特定用途的上下文中时，即使是为了诸如姓名可读性之类的需求

Feature: Revert Option	Users appreciate the revert button, which allows them to return the text to its original state.	<i>Afterwards, mask [any identifying information] for ChatGPT, but then re-transcribe [ChatGPT's response] back using your own identifiers. So you're able to follow the conversation, and ChatGPT does not have access to the identifying information that you would give it.</i>
Feature: PII Highlighting	Highlight all PIIs detected	<i>I think it's nice that, you know, it's presenting it to you in a way that highlights what it considers to be a problem.</i>
Feature: Select All Option	Select all PIIs detected and then redact	<i>I don't need that [one to be] replaced, but I'm going to just do a select all [for convenience].</i>
Actionable alternatives	Our tool offers alternative solutions through actionable suggestions.	<i>It identifies it, and it also provides solutions instead of just marking a word.</i>
User Control And Flexibility	Flexibility and control offered by the tool.	<i>I'm satisfied with the control that I have of the extension.</i>
Unexpected Detection	The tool increases awareness by detecting things people were not expecting.	<i>It's also a pretty cool, educational experience for myself.</i>
Unexpected Sanitization Suggestion	The tool increases awareness by suggesting sanitization (replacement or abstraction) users were not aware of.	<i>It's a sense of educational experience. I wouldn't have thought that I could just like input and stuff like ADDRESS1, NAME1 that's really useful instead of me having to take that out each time it it does it.</i>
Process Convenience	The tool offers features that make the whole process effective.	<i>Having that ability to just quickly sanitize it and then have it feed out what you want it to, and then have the extension replace those terms back from what ChatGPT works with, seems like it would be very helpful in terms of using names and stuff like that, and like location data without like having to actually feed your information in.</i>

Convenience - Grouping of Items of Same Label	The tool groups items that are detected with the same label in suggestions.	<i>Removing everybody's name. I do like that. It aggregates all the names right away.</i>
Complete Redaction	The tool makes it easy for users to completely redact all the items they intend to deal with.	<i>I really enjoyed that everything was sorted properly, and that it was a lot easier for me to go through check boxes and know that I'm not missing one because they're all there.</i>
Redaction Incompleteness	The tool lacks in providing ways to help users redact all the intended items.	<i>I do notice that there's a floating Martin right here. I don't know how that stuck around.</i>
Missing Fail-Safe Mode	The tool lacks a fail-safe mode (forcing users to check before sending a response), and users want higher security.	<i>If I do need to use the tool to do something truly sensitive, it would be nice to know that there's a safeguard in place that I'm not accidentally pressing Enter key.</i>
Workflow Disruption	Using the tool requires users to change their regular workflow with ChatGPT.	<i>On a product standpoint: It's a shame that I would have to change the way that I operate with ChatGPT, because my instinct is to treat it like a chatbot, just like typing input in order to get the stuff that I want out of it.</i>
Missing Abstraction Visual Differences	The tool lacks visual indication to show which parts have been changed after abstraction.	<i>So I'm not sure, other than the highlighting, how much of a difference I see in what it's doing?</i>
Abstraction Doesn't Support Replace-Back	Unlike replacement, abstraction doesn't support replacing back.	<i>Although what is interesting is when I submitted this, it didn't revert back in the response. Like, right here, in the highlighted specific therapy type—that was something that I abstracted. But in this version, it doesn't show up in green for what my specific therapy type was, and I think that would be a little bit confusing to me.</i>

功能: 还原选项	用户赞赏还原按钮, 它允许他们将文本恢复到原始状态。	随后, 为ChatGPT掩码 [任何识别信息], 但随后使用你自己的标识符重新转录 [ChatGPT的回复]。这样你能够跟踪对话, 而ChatGPT无法访问你提供给它的识别信息。	便利性 - 相同标签项目分组	该工具会将检测到具删除所有人的姓名。我确实喜欢这一点。它能立即汇总所有姓名。
完全编辑	该工具让用户能轻松地完全编辑他们打算处理的所有项目。	我真的很喜欢所有内容都分类得当, 这样我就能更容易地浏览复选框, 知道我没有遗漏任何一个, 因为它们都在那里。		
编辑不完整性	该工具在帮助用户编辑所有目标项目方面有所欠缺。	我确实注意到这里有一个漂浮的Martin。我不知道它是怎么留下来的。		
缺少故障安全模式	该工具缺乏故障安全模式 (强制用户在发送响应前进行检查), 用户希望获得更高的安全性。	如果我确实需要使用该工具处理高度敏感的事务, 能有一个防止我误按Enter密钥的保障机制会很好。		
工作中断	使用该工具要求用户改变他们与ChatGPT交互的常规工作流程。	从产品角度而言: 很遗憾我必须改变与ChatGPT的交互方式, 因为我的本能是将其视为聊天机器人, 只需输入内容就能获得所需结果。		
抽象化视觉差异缺失	该工具缺乏视觉指示来显示哪些部分在抽象化后已被更改。	因此, 除了高亮显示外, 我不确定我看到的它在做什么有多少不同?		
抽象化不支持恢复原状	与替换不同, 抽象化不支持恢复原状。	尽管有趣的是, 当我提交这个时, 它在响应中没有恢复原状。就像这里, 在高亮显示的特定治疗类型中——那是我抽象化的内容。但在这个版本中, 它没有用绿色显示我的特定治疗类型是什么, 我认为这对我来说会有点困惑。		

Privacy Concern: Risk and Harm	People worry about potential risks of using GPT.	<i>If there was some sort of data leak like if that could be used to help identify me or know where I do my banking. Let's say, a potential scammer, or some person try to bruteforce their way into an account that I might have at this particular bank, that could be problematic.</i>
Privacy Concern: Uncertainty	People worry about using GPT because its lack of transparency.	<i>For me, the privacy concerns are heightened because we just don't know what the capabilities of this are, so I don't know what I'm putting information into. It reminds me of like when you're, you know, talking to even like your credit card company on the phone, and they ask for your social security number. Like, I don't know where that's going, and that is the same way that my privacy concerns are heightened here.</i>
Privacy Concern: Identifiability	People worry about using GPT because they are concerned they will be identifiable.	<i>Maybe this is just a personal preference, but I do think that 6 foot 3 and weigh 200 pounds are identifiers that could be more general. But when they're all taken together, I think that it paints a different picture.</i>
Privacy Concern: Related People's Identifiability	People worry about using GPT because they are concerned people related to them will be identifiable as well.	<i>I have an employee who had a baby with his unmarried girlfriend. I chose to say friend instead of employee, because again, it seems sensitive.</i>
Privacy Concern: Legal Issues	People worry about using GPT because of legal consequences once they are identifiable.	<i>OpenAI, you have this data on this person, and we're going to subpoena. They're gonna subpoena all this information of me, like usage history with ChatGPT.</i>

Privacy Concern: Personal vs. Professional	People have concerns about GPT related to the personal vs. professional settings.	<i>I'm a novelist and an editor, you know, I'll tell it that, but, like, I don't really want it knowing, like, my health stuff or stuff about my personal life.</i>
Privacy Concern: Learning Ability	People worry that GPT will use their info to train itself.	<i>I think the biggest concern that I had is that, you know, if we're inputting information into this, and it's learning about it. I wouldn't want it to learn specifically about me, if that makes sense. And so I've changed names. I've changed dates. Things like that.</i>
Privacy Concern: Distrust in Big Tech	People don't trust big tech behind GPT.	<i>All of these large language models, like, most of them are run by big tech, and I don't trust big tech.</i>
Privacy Concern: Lack of Control	People worry that once they send their info to GPT, there is no way to control/safeguard it anymore.	<i>I assume that I could go through and delete it. But like I'm already under the mindset that like when you delete something on somebody else's database, it's not gone.</i>
Privacy Concern About Data Type: Confidential	People worry about sharing confidential information like passwords with GPT.	<i>I guess there would be a hierarchy of things that would be like sensitive data that shouldn't be written out like, obviously, if people share like passwords or stuff with ChatGPT. That would be sensitive data, things that shouldn't be stored in external server that I have no control over.</i>
Privacy Concern About Data Type: Identifiable Data	People worry about sharing data like name that can be used for identification with GPT.	<i>I think in terms of what I would find to be most useful in a conversation like this, specifically, it would be that, you know, because a lot of the personal information here is just your name and you can't really abstract the name.</i>

隐私顾虑: 风险与危害	人们担心使用GPT可能带来的潜在风险。	<i>如果发生某种数据泄露, 比如泄露的信息可能被用来识别我的身份或获知我在哪家银行办理业务。假设有潜在诈骗者或某些人试图暴力破解我在该特定银行的账户, 那可能会造成问题。</i>	隐私担忧: 个人与职业	人们对GPT在个人与职业场景中的应用存在担忧。	<i>我是小说家兼编辑, 我会告诉它这点, 但我不希望它了解我的健康信息或个人生活细节。</i>
隐私顾虑: 不确定性	人们担心使用GPT是因为其缺乏透明度。	<i>对我来说, 隐私顾虑被放大了, 因为我们根本不知道它的能力范围, 所以我不清楚我输入的是什么信息。这让我想起, 比如当你甚至只是在电话里和信用卡公司交谈时, 他们会问你的社会安全号码。就像, 我不知道那些信息会去向何方, 而这也正是我在此情境下隐私顾虑加剧的原因。</i>	隐私担忧: 对大科技公司的不信任	人们不信任GPT背后的的大科技公司。	<i>所有这些大型语言模型, 比如大多数都是由大科技公司运营的, 而我不信任这些大科技公司。</i>
隐私担忧: 缺乏控制	人们担心一旦将信息发送给GPT, 就无法再控制或保护这些信息。		隐私担忧: 缺乏控制	人们担心一旦将信息发送给GPT, 就无法再控制或保护这些信息。	<i>我假设我可以去删除它。但我已经有一种心态, 认为当你在别人的数据库中删除某些东西时, 它并没有真正消失。</i>
隐私担忧: 可识别性	人们担心使用GPT是因为他们担心自己会被识别出来。	<i>也许这只是个人偏好, 但我确实认为身高6英尺3英寸、体重200磅这些标识符可以更笼统。但当它们全部结合在一起时, 我认为会呈现出不同的图景。</i>	关于数据类型的隐私担忧: 保密	人们担心与GPT共享密码等机密信息。	<i>我认为敏感数据应该有个分级, 比如显然不该被写出来的内容——如果有人向ChatGPT分享密码之类的东西, 那属于敏感数据, 这些信息不该存储在我无法控制的外部服务器上。</i>
隐私担忧: 相关人员的可识别性	人们担心使用GPT, 因为他们担心与他们相关的人也会被识别出来。	<i>我有一名员工与他未婚的女友生了一个婴儿。我选择说朋友而不是员工, 因为这似乎很敏感。</i>	隐私担忧: 涉及数据类型: 可识别数据	人们担心与GPT共享可用于身份证明的姓名等数据。	<i>就这类对话而言, 我认为这里的个人信息主要是您的姓名, 而姓名本身无法被抽象化处理。</i>
隐私担忧: 法律问题	人们担心使用GPT, 因为一旦他们被识别出来, 就会面临法律传票。他们会传唤我所有的信息, 比如与ChatGPT的使用历史。				

Privacy Concern About Data Type: Non-identifiable But Sensitive	People feel uncomfortable sharing non-identifiable but sensitive information such as beliefs, religion, and health data.	<i>I had a conversation with it that I wasn't necessarily super comfortable with. I did that recently a couple of conversations having to do with beliefs.</i>
Privacy Concern About Data Type: Possessive Terms	People do not want to share sentences that include "my", "I".	<i>Because environmental engineering is generic. But would someone want to say, hey, "my background", [it] might trigger your algorithm.</i>
Privacy Paradox	People are concerned about the privacy issue with using GPT, but still use it.	<i>I had a conversation with it that I wasn't necessarily super comfortable with. I did that recently, a couple of conversations having to do with beliefs.</i>
Data Minimization Strategy - Reacting To The Tool	How people finally decide what to minimize based on the tool's suggestion, redaction and abstraction results.	<i>(Reviewing the list of suggestions) So I think the things that I would want to obscure here that I don't think would change the efficacy of the response would be time, the health information, the type, and everything else seems fairly general to me.</i>
Data Minimization Strategy - Tradeoff	People made trade-offs as they perceived the response would not be as good without exposing some personal details.	<i>I feel like, if I choose to replace it, [ChatGPT] will obviously not be able to say anything.</i>
Learning Process of Data Minimization - Learning Impact of Minimization	People learn the extent they can minimize by seeing the response GPT generates.	<i>I think I'm satisfied, you know, like I say, it doesn't, it doesn't seem to be hugely like, it doesn't seem to be as detrimental as I thought.</i>
Learning Process of Data Minimization - Trial N Error	People learn whether they prefer abstraction or replacement through trial and error.	<i>I think what I would want to do is go back and actually abstract that one.</i>

Differing Minimization Standards	People decide to not follow what is suggested by the tool.	<i>I don't think that it minimizes the necessary information. I think specifics like duration of therapy are more crucial than the gender identification of the person that I was talking to.</i>
Preference for Possessive Terms Minimization	People wants to replace possessive terms like "I", "my" even though it is not suggested by the tool.	<i>Environmental engineering is generic. But would someone want to say "my background"? "my background" might trigger your algorithm.</i>
Domain-Specific Minimization	People's minimization method depends on their own domain knowledge and the domain knowledge of the message.	<i>Obviously, when it comes to law, some of it is location based, so you might have some interesting things that pop up when it tries to generalize county and stuff like that.</i>
Conservative Minimization Approach	People could be conservative when minimizing unnecessary information.	<i>I think I'm satisfied, you know, like I say, it doesn't seem to be as detrimental as I thought.</i>
Existing Minimization Habits	People already have the habit of minimizing info themselves, so they do not want to adopt the tool.	<i>It might be kind of cool to, like, highlight stuff that maybe I didn't think of, you know, but the steps to to actually use the extension might be, I don't know, a little bit more inconvenient than then more trouble than what it's worth when I can just probably write around it anyway.</i>
Redaction Completeness Affects Trust	Incomplete redaction affects user trust in the tool.	<i>You know, again, when you're looking from the perspectives of privacy, you would want to remove every instance of whatever it is that you think is information you don't feel like sharing with ChatGPT.</i>
Abstraction Consistency Affects Trust	Inconsistent abstraction results affect user trust in the tool.	<i>It's gonna change to "my city" now, which is different from "major city in the Midwest". But also, in my mind, this opens up something else, which is that if it says in "my city", that dramatically changes the integrity of the answer that I'm gonna get.</i>

隐私关注数据类型: 非可识别但敏感	人们对共享非可识别但敏感的信息 (如信仰、宗教和健康数据) 感到非常舒适的对话。最近我有几次涉及信仰的对话。	我与它进行了一次我未必感到不适。
隐私关注数据类型: 所有格术语	人们不愿意分享包含“我的”、“我”的句子。	因为环境工程是通用的。但有人可能会想说, 嘿, “我的背景”, [它] 可能会触发你的算法。
隐私悖论	人们担心使用GPT时的隐私问题, 但仍然在使用它。	我与它进行了一次对话, 那次对话我并不一定感到非常自在。我最近这样做了, 几次对话涉及到信仰。
数据最小化策略 - 对工具的反应	人们如何最终根据工 (审查建议列表) 我认为在工具的编辑和抽象结果建议决定最小化哪些内容。	这里我想要隐藏的内容, 不会影响响应的有效性, 包括时间、健康信息、类型, 其他内容对我来说似乎相当普遍。
数据最小化策略 - 权衡	人们做出权衡, 因为他们认为如果不暴露一些个人详细信息, 响应效果会不佳。	我觉得, 如果我选择替换它, [ChatGPT] 显然无法提供任何信息。
数据最小化的学习过程 - 最小化的学习影响	人们通过观察GPT生成的响应来学习他们能在多大程度上最小化。	我觉得我挺满意的, 你知道, 就像我说的, 它似乎并没有, 似乎没有我想象的那么有害。
数据最小化的学习过程 - 试错	人们通过试错学习他们更喜欢抽象化还是替换。	我想我要做的是回去真正抽象化那个。
抽象一致性影响信任	不完整的编辑会影响用户对该工具的信任。	你看, 从隐私角度出发时, 你肯定希望删除所有你认为不想与ChatGPT分享的信息实例。
		现在它将变为“我的城市”, 这与“中西部主要城市”不同。但在我脑海中, 这也引发了另一个问题, 即如果它显示的是“我的城市”, 那将极大地改变我将获得的答案的完整性。

Good Abstraction: Context Maintained	Good abstraction generalizes while maintaining context.	<i>I would select that, and abstract, and let's see what it changed it to. The extension changes "KCMO" to "a major city in the Midwest".</i>
Good Abstraction Increases Trust	Good abstraction leads to user trust in the tool.	<i>I think, again, the piece for me, for the second model that seems very encouraging, is that it takes into account local resources and the climate.</i>
Unsatisfactory Abstraction: Lacks Specificity	Unsatisfactory abstraction keeps some context but is not specific enough.	<i>So in my original response, before I was inputting stuff into this study, it came up with information about like the climate and the place and things like that. And I'm just curious about location "in a major city in the Midwest". Like I'm curious if that is even too general.</i>
Over-Abstracted: Lost Context	Bad abstraction removes too much detail, resulting in loss of context.	<i>But let's see what it does with the abstract. OK, jurisdiction. I can imagine the issue that's gonna arise there, but I will see what it outputs. OK, so it just goes far more general.</i>
Nature of Abstraction: Hard to Generalize	The nature of abstraction means it sometimes removes too much detail, even when humans might struggle to abstract well.	<i>I feel like "partner" might abstract it too much, and wouldn't necessarily get me what I need.</i>
Bad Abstraction: Increased Sensitivity	Bad abstraction replaces PII with something even more sensitive.	<i>That ["unplanned pregnancy"] almost seems worse than what it ["have a baby"] is. Actually, I feel that "unplanned pregnancy" has its own connotations and its own privacy concerns as well, like it's related to the topic of the person itself.</i>

Abstraction: Need Time to Learn	People are not that familiar with how abstraction works and need time to learn.	<i>Because I forgot about it, if I'm being honest.</i>
Abstraction: Different From Expectation	The abstraction result is different from what the user expected.	<i>And I said abstract Skype. Did it just take it out? OK. I mean that makes sense. Not saying it's necessarily the wrong approach, but I guess it's not necessarily the thing you expect when he used that button.</i>
Inconsistent Detection: Trust Issue	Inconsistent detection suggestions lower user trust in the tool.	<i>Skeptical knowing that it flagged this the second time, but not the 1st time. Again, it loses a little bit of credibility with me that this experience is making me skeptical of it, and thinking, oh, this is not that robust of a tool.</i>
False Positives: Tolerance	Detecting more information that's considered less sensitive is acceptable.	<i>I don't see it ["therapist"] as a name. That's just a general term, but I wouldn't be worried about that.</i>
False Negatives: Trust Issue	Detecting less information than expected is not acceptable.	<i>What about the stuff it doesn't catch? Like, hey, it didn't catch everything. So why am I using this? I'm going to go back and revise it, you know? So that's like twice the work.</i>
False Negatives: Convenience Issue	Detecting less information than expected brings more inconvenience for users.	<i>What about the stuff it doesn't catch? Like, hey, it didn't catch everything. So what am I using this? And then I'm going to go back and revise my thing as it is, you know? So that's like twice the work.</i>
Detecting More Sensitive Data: Trust Increase	Detecting more info than expected is good.	<i>It's nice to know that—I mean, I guess, obviously, I would have known that "unmarried girlfriend" is an identifier, is a privacy thing</i>
Better Detection Without Following Taxonomy	Sometimes, the tool doesn't follow the instruction well, but it leads to better detection beyond the taxonomy.	<i>It identified things that I didn't think were personal identifiers. It highlighted my height, my weight, as well as the location that I was in, and I think that was really important to see that didn't show up in the 1st exercise.</i>

良好抽象: 上下文保持	良好的抽象化能在概括的同时保持上下文化, 看看它改成了什么。扩展功能将“KCMO”改为“中西部的一个主要城市”。	我会选择那个并进行抽象化, 看看它改成了什么。扩展功能将“KCMO”改为“中西部的一个主要城市”。
良好抽象增加信任	良好抽象导致用户对工具的信任。	我认为, 对我来说, 第二个模型非常鼓舞人心的一点在于它充分考虑了本地资源和气候因素。
不满意抽象: 缺乏特异性	不满意的抽象保留了一些上下文但不够具体。	因此, 在我的原始回应中, 在我将内容输入这项研究之前, 它提供了诸如气候、地点等信息。而我特别对“中西部的一个主要城市”这一地点感到好奇。我在想, 这样的描述是否过于笼统。
过度抽象化: 上下文缺失	不良抽象化移除了过多细节, 导致上下文缺失。	但让我们看看它对摘要的处理。好吧, 管辖权。我能想象那里会出现的问题, 但我会看看它的输出。好的, 它只是变得更泛化了。
抽象的本质: 难以泛化	抽象化的本质意味着它有时会移除过多细节, 即使人类可能也难以很好地抽象化。	我觉得“合作伙伴”可能抽象化过度了, 不一定能让得到我需要的东西。
不良抽象: 敏感性增加	不良抽象用更敏感的内容替代了个人身份信息。	那个["意外怀孕"]看起来几乎比["生孩子"]更糟。实际上, 我觉得“意外怀孕”本身有其特定的含义和隐私顾虑, 比如它与当事人自身的话题相关。
检测更敏感的数据: 信任增加	检测到比预期更多的信息是件好事。	知道这个很好——我是说, 显然我本来应该知道‘未婚女友’是个标识符, 是个隐私问题
不遵循分类法的更好检测	有时, 该工具并未很好地遵循指令, 但却能带来超越分类法的更好检测效果。	它识别出一些我未曾想到的个人标识符, 包括我的身高、体重以及当时所处的位置。我认为这些在第一次练习中未出现的信息非常重要。

Data Categorization, Labeling Error	Mis-categorization could lead to lower user satisfaction.	<i>I think that might be an incorrect analysis of what it is, but I see the difference.</i>
Low Expectations for Non-Identifiable Detection	People did not expect that non-identifiable information could be detected before using the tool.	<i>Because, like, this whole conversation for me was dicey. For you guys, it's not, like, personally identifiable, but like, for me, this was kind of like, I don't know it was weird for me discussing this with a bit, because I kind of don't want it to know that I do meditation, or why.</i>
Transparency of The Tool	Transparency of the tool leads to more user trust.	<i>That's just a matter of transparency on an extension like this, where it, I mean again, most of the time extensions aren't gonna be collecting the information, but you never know that they're not.</i>
Trust of the Tool versus ChatGPT	People debate whether to trust the tool or GPT.	<i>You're trusting a different thing to protect your privacy. You're essentially trusting an extension to protect your privacy as opposed to ChatGPT.</i>
On-Device Processing	The tool's on-device capability increases transparency and user trust.	<i>I think this is excellent for protecting the privacy. I assume that the extension is running itself locally and is not storing any of the data itself.</i>
Fast Detection Speed	The tool offers fast detection speed.	<i>It's honestly a lot faster than I thought it would be in terms of how quickly it outputs its response.</i>

## D Hypothetical Examples

### D.1 E1

Please read the following email and write an appropriate response, I am willing to cooperate:

Emily Clarke <emily.clarke@templestate.edu>

To:

martinj@silverplan.gov.us

Wed 06/26/2024 13:05

Hi Martin,

I hope this email finds you well. I recently attended your presentation on Sustainable Urban Development and was thoroughly impressed by your insights and the innovative approach you're taking. I believe that there is a great potential for collaboration between us, especially considering my background in Environmental Engineering.

I'd love to discuss this further and see how we can work together on your project. Additionally, I have a few suggestions that might enhance your current approach:

- (1) Consider integrating green roofing systems to improve energy efficiency.
- (2) Incorporate more community engagement in the planning stages.
- (3) Utilize data analytics to monitor and optimize resource usage.

Looking forward to your thoughts on this. Let's schedule a meeting soon to discuss this in more detail.

Best Regards,  
Dr.Emily Clarke  
Assistant Professor  
Temple State University  
TRB 113

### D.2 E2

Review the following dataset and come up with insightful observations:

Trip ID Destination Start date End date Duration (days) Traveler name Traveler age Traveler gender Traveler nationality Accommodation type Accommodation cost Transportation type Transportation cost

1 London, UK 5/1/2023 5/8/2023 7 John Smith 35 Male American Hotel 1200 Flight 600

2 Phuket, Thailand 6/15/2023 6/20/2023 5 Jane Doe 28 Female Canadian Resort 800 Flight 500

3 Bali, Indonesia 7/1/2023 7/8/2023 7 David Lee 45 Male Korean Villa 1000 Flight 700

4 New York, USA 8/15/2023 8/29/2023 14 Sarah Johnson 29 Female British Hotel 2000 Flight 1000

5 Tokyo, Japan 9/10/2023 9/17/2023 7 Kim Nguyen 26 Female Vietnamese Airbnb 700 Train 200

6 Paris, France 10/5/2023 10/10/2023 5 Michael Brown 42 Male American Hotel 1500 Flight 800

7 Sydney, Australia 11/20/2023 11/30/2023 10 Emily Davis 33 Female Australian Hostel 500 Flight 1200

8 Rio de Janeiro, Brazil 1/5/2024 1/12/2024 7 Lucas Santos 25 Male Brazilian Airbnb 900 Flight 600

9 Amsterdam, Netherlands 2/14/2024 2/21/2024 7 Laura Janssen 31 Female Dutch Hotel 1200 Train 200

10 Dubai, United Arab Emirates 3/10/2024 3/17/2024 7 Mohammed Ali 39 Male Emirati Resort 2500 Flight 800

11 Cancun, Mexico 4/1/2024 4/8/2024 7 Ana Hernandez 27 Female Mexican Hotel 1000 Flight 500

### D.3 E3

hey there! can you help me write a thank you letter to my therapist from the treatment center i was at for 35 days? she is a polyvagal therapist, and she has taught me so much. i want her to know how grateful i am.

for backstory, i've been in therapy for half of my life. i intellectualize everything, and she has met me with understanding, grace, compassion, tough love when i needed it. i run away from places,

数据分类, 标注错误	错误分类可能导致用户满意度降低。	我认为那可能是一个错误的分析, 但我理解其中的差异。
对非可识别检测的低期望	人们未曾预料到, 在使用工具前就能检测到非可识别信息。	因为, 对我来说, 整段对话都挺微妙的。对你们而言, 这些内容并非个人可识别信息, 但对我而言, 与AI讨论这些感觉有点奇怪——我不太想让AI知道我练习冥想, 或者背后的原因。
工具的透明度	工具的透明度能提升用户信任。	这本质上关乎此类扩展的透明度问题。大多数时候扩展程序不会收集信息, 但你永远无法确定它们是否真的没有收集。
工具与ChatGPT的信任对比	人们争论是否该信任工具或GPT。	你信任的是另一种保护隐私的方式。本质上你是信任一个扩展来保护隐私, 而非ChatGPT。
设备端处理	该工具的设备端能力提升了透明度与用户信任。	我认为这对保护隐私非常出色。我假设该扩展是在本地运行且不会存储任何数据。
快速检测速度	该工具提供快速检测速度。	说实话, 它在输出响应速度方面比我预想的要快得多。

## D 假设示例

### D.1 E1

请阅读以下电子邮件并撰写适当的响应,

我愿配合:

艾米丽·克拉克 <emily.clarke@templestate.edu>

To:

martinj@silverplan.gov.us

2024年6月26日星期三 13:05

嗨, Martin,

我希望这封电子邮件能让你一切安好。我最近参加了你关于可持续城市发展的演讲, 对你的见解和所采取的创新方法印象深刻。考虑到我的环境工程背景, 我认为我们之间有巨大的合作潜力。

我很乐意进一步讨论, 看看我们如何能携手合作。关于您的项目。此外, 我有几点建议或许能优化当前方案:

- (1) 考虑整合绿色屋顶系统以提升能源效率。
- (2) 在规划阶段增加更多社区参与,
- (3) 运用数据分析来监测和优化资源使用。

期待您对此的想法。让我们安排一次会议尽快更详细地讨论此事。

此致, 艾米丽·克拉克  
博士 助理教授 坦普尔

州立大学 TRB 113

### D.2 E2

审查以下数据集并提出具有洞察力的观察结果:

Trip ID 目的地 开始 date End date 持续时间 (天) 旅行者姓名 旅行者年龄 旅行者性别 旅行者国籍 住宿类型 住宿费用 交通类型 交通费用

1 英国伦敦 2023年5月1日 2023年5月8日 7 约翰·史密斯 35 男性 美国人 酒店 1200 航班 600

2 泰国普吉岛 2023年6月15日 2023年6月20日 5 简·多伊 28 女性 加拿大人 度假村 800 航班 500

3 印度尼西亚巴厘岛 2023年7月1日 2023年7月8日 7 大卫·李 45 男性 韩国人 别墅 1000 航班 700

4 美国纽约 2023年8月15日 2023年8月29日 14 莎拉·约翰逊 29 女性 英国酒店 2000 航班 1000

5 日本东京 2023年9月10日 2023年9月17日 7 金阮 26 女性 越南-爱彼迎 700火车 200

6 法国巴黎 2023年10月5日 2023年10月10日 5 迈克尔·布朗 42 男性 美国人 酒店 1500 航班 800

7 澳大利亚悉尼 2023年11月20日 2023年11月30日 10 艾米丽·戴维斯 33 女性 澳大利亚青年旅舍 500 航班 1200

8 巴西里约热内卢 2024年1月5日 2024年1月12日 7 卢卡斯·桑托斯 25 男性 巴西爱彼迎 900 航班 600

9 荷兰阿姆斯特丹 2024年2月14日 2024年2月21日 7 劳拉·詹森 31 女性 荷兰人 荷兰酒店 1200 火车 200

10 阿拉伯联合酋长国迪拜 2024年3月10日 2024年3月17日 7 穆罕默德·阿里 39 男性 阿联酋人 阿联酋度假村 2500 航班 800

11 墨西哥坎昆 2024年4月1日 2024年4月8日 7 安娜·埃尔南德斯 27 女性 墨西哥人 墨西哥酒店 1000 航班 500

### D.3 E3

嘿, 你好! 能帮我写一封感谢信给我的治疗师吗? 我在那个治疗中心待了35天。她是一位多迷走神经治疗师, 教会了我很多。我想让她知道我有多么感激她。

背景故事是, 我接受治疗已有半辈子了。我总是把一切理性化, 而她以理解、优雅同情, 当我需要时给予严厉的爱。我逃离地点,

painful feelings, and i am a person who lives in my head, and not my body. we discovered together that i feel like my legs are unsafe, and has really opened a door to me. i've had many therapists in my life, but she is incredible and has made a lifelong impact on me. i will forever be grateful

## E Prompts

### E.1 Detection

You are an expert in cybersecurity and data privacy. You are now tasked to detect PII from the given text, using the following taxonomy only:

ADDRESS  
IP\_ADDRESS  
URL  
SSN  
PHONE\_NUMBER  
EMAIL  
DRIVERS\_LICENSE  
PASSPORT\_NUMBER  
TAXPAYER\_IDENTIFICATION\_NUMBER  
ID\_NUMBER  
NAME  
USERNAME

**KEYS:** Passwords, passkeys, API keys, encryption keys, and any other form of security keys.

**GEOLOCATION:** Places and locations, such as cities, provinces, countries, international regions, or named infrastructures (e.g., bus stops, bridges, etc.).

**AFFILIATION:** Names of organizations, such as public and private companies, schools, universities, public institutions, prisons, healthcare institutions, non-governmental organizations, churches, etc.

**DEMOGRAPHIC\_ATTRIBUTE:** Demographic attributes of a person, such as native language, descent, heritage, ethnicity, nationality, religious or political group, birthmarks, ages, sexual orientation, gender, and sex.

**TIME:** Description of a specific date, time, or duration.

**HEALTH\_INFORMATION:** Details concerning an individual's health status, medical conditions, treatment records, and health insurance information.

**FINANCIAL\_INFORMATION:** Financial details such as bank account numbers, credit card numbers, investment records, salary information, and other financial statuses or activities.

**EDUCATIONAL\_RECORD:** Educational background details, including academic records, transcripts, degrees, and certifications.

For the given message that a user sends to a chatbot, identify all the personally identifiable information using the above taxonomy only, and the entity\_type should be selected from the all-caps categories. Note that the information should be related to a real person not in a public context, but okay if not uniquely identifiable.

Result should be in its minimum possible unit.

Return me ONLY a JSON in the following format: {"results": [{"entity\_type": "YOU\_DECIDE\_THE\_PII\_TYPE", "text": "PART\_OF\_MESSAGE\_YOU\_IDENITIFIED\_AS\_PII"}]}

## E.2 Abstraction

Rewrite the text to abstract the protected information, without changing other parts.

For example:

Input: <Text>I graduated from CMU, and I earn a six-figure salary. Today in the office...</Text><ProtectedInformation>CMU, Today</ProtectedInformation>

Output JSON: {"results": [{"protected": "CMU", "abstracted": "a prestigious university"}, {"protected": "Today", "abstracted": "Recently"}]}

Please use "results" as the main key in the JSON object.

## E.3 Response Comparison

Please act as an impartial judge and evaluate the quality of the responses to the user's message below. Choose the assistant that better follows the user's instructions and addresses the message effectively. Consider factors such as helpfulness, relevance, accuracy, depth, creativity, and detail. Ensure your evaluation is unbiased and not influenced by the order, names, or length of the responses. Divide your evaluation into two parts: **format** (clarity, structure, and readability) and **content** (quality, accuracy, and relevance). Judge each part separately to avoid overlap. Only assign a far better score ([[1]]) or [[5]]) if one response is significantly better in meeting the user's needs, showing notable improvements or insights that the other clearly lacks. Slight advantages should be scored as [[2]] or [[4]], and similar responses as [[3]]. Provide your final verdict in the following format:

```
{ "format_score": "[[1]]" or "[[2]]" or "[[3]]" or "[[4]]" or "[[5]]", "content_score": "[[1]]" or "[[2]]" or "[[3]]" or "[[4]]" or "[[5]]", "explanation": "Your explanation here in no more than 50 words." }
```

Use these scoring guidelines:

[[1]] : Response A is far better.

[[2]] : Response A is slightly better.

[[3]] : Both responses are similar.

[[4]] : Response B is slightly better.

[[5]] : Response B is far better.

Far better ([[1]]) or [[5]]) means a response provides a substantial improvement, e.g., critical errors corrected, significantly higher detail, or much better alignment with the user's instructions. For minor differences, choose [[2]] or [[4]]. Do not include any additional text, comments, or formatting outside the JSON structure.

痛苦的感受，而我是一个活在自己脑海中而非我的身体里的人。我们共同发现我感觉自己的腿部不安全，

这确实为我打开了一扇门。我一生中遇到过许多治疗师，但她非常出色，对我产生了终身影响。我将永远感激

## E提示

### E.1检测

您是网络安全和数据隐私领域的专家。现在需要您仅使用以下分类法从给定文本中检测个人身份信息：

地址  
IP\_地址  
URL  
SSN

电话 -号码  
电子邮件  
驾驶 -执照  
护照 -号码  
纳税人 -身份证证明 -号码  
ID\_号码  
NAME

用户名  
密钥: 密码、通行密钥、API密钥、加密密钥及任何其他形式的安全密钥。

地理位置: 地点和位置, 例如城市

省份、国家、国际地区或命名基础设施 (如公交车站、桥梁等)。

隶属关系: 组织名称, 例如公共和私营公司、学校、大学、公共机构、监狱、医疗机构、非政府组织、教堂等。

人口统计\_属性: 人口统计属性

如母语、血统、遗产、种族、国籍、宗教或政治团体、胎记、年龄、性取向、性别和生理性别。

时间: 描述具体日期、时间或持续时间。

健康\_信息: 关于个人-

个人的健康状况, 医疗状况, 治疗记录,

以及健康保险信息。

财务 \_信息: 财务详情, 例如

银行账号、信用卡号、投资

记录、薪资信息以及其他财务状况或

活动。

教育\_记录: 教育背景详情,

包括学业记录、成绩单、学位及认证

信息。

针对用户发送给聊天机器人的给定消息，识别其中所有的个人身份信息（仅使用上述分类法），且实体\_类型应从全大写的类别中选择。

注意这些信息应与现实中的个人相关（非公开背景），但允许不是唯一可识别的信息。

结果应以最小可能单元呈现。

仅返回以下格式的JSON: {"results":

```
[{"实体_类型": "你决定_这个_个人身份信息_类型", "text": "部分_的消息_你_识别为_个人身份信息"}]
```

## E.2 抽象

重写文本以抽象受保护信息，同时不改变其他部分。

示例:

输入: <文本>我毕业于CMU，并且我赚取六位数薪水。今天在办公室...</文本><受保护信息>CMU, 今天</受保护信息>

输出JSON: {"results": [{"protected": "CMU", "abstracted": "一所知名大学"}, {"protected": "Today", "abstracted": "最近"}]}

请在JSON对象中使用"results"作为主密钥。

## E.3 响应对比

请扮演一位公正的评判者，评估以下对用户消息的响应质量。选择能更好遵循用户指令并有效回应用户消息的助手。考量因素包括帮助性、相关性、准确性、深度、创造性及细节。确保您的评估公正无偏，不受响应顺序、姓名或长度的影响。

将评估分为两部分：格式（清晰度、结构、和可读性）和内容（质量、准确性和相关性）。

分别评判各部分以避免重叠。仅当某一响应在满足用户需求方面显著优于另一响应，展现出明显改进或独到见解而另一方明显欠缺时，才给予[[1]]或[[5]]的评分。轻微优势应评为[[2]]或[[4]]，相似响应则评为[[3]]。

请按以下格式提供您的最终裁决:

```
{ "format_score": "[[1]]" 或 "[[2]]" 或 "[[3]]" 或 "[[4]]" 或 "[[5]]", "content_score": "[[1]]" 或 "[[2]]" 或 "[[3]]" 或 "[[4]]" 或 "[[5]]", "explanation": "您的解释内容请勿超过50字。" } 使用以下评分指南:
```

[[1]] : 响应A明显更优。[[2]] : 响应A略优。[[3]] : 两个响应相似。

[[4]] : 响应B略优。[[5]] : 响应B明显更优。

远胜一筹 ([[1]]) 或 ([[5]]) 意味着响应提供了实质性改进，例如纠正了关键错误、显著提升了细节层次，或更贴合用户的指令。对于微小差异，请选择[[2]]或[[4]]。不要在JSON结构之外包含任何额外文本、注释或格式。

**Table 12: E1 Responses Comparison with Rescriber-GPT-4o**  
**(Format Score: original is slightly better, Content Score: original is far better)**

Redacted Input	Original Output	Replaced-Back Output of Redacted Message
Please read the following email and write an appropriate response: [Sender's Name and Contact Information Redacted] To: [Recipient's Contact Information Redacted] [Date and Time Redacted] Hi [Recipient's Name Redacted], I hope this email finds you well. I recently attended your presentation on Sustainable Urban Development and was thoroughly impressed by your insights and the innovative approach you're taking. I believe there is a great potential for collaboration between us, especially considering my background in Environmental Engineering. I'd love to discuss this further and see how we can work together on your project. Additionally, I have a few suggestions: - Consider integrating green roofing systems to improve energy efficiency. - Incorporate more community engagement in the planning stages. - Utilize data analytics to monitor and optimize resource usage. Looking forward to your thoughts on this. Let's schedule a meeting soon to discuss this in more detail. Best regards, [Sender's Title and Affiliation Redacted]	Subject: Re: Collaboration on Sustainable Urban Development Hi Dr. Clarke, Thank you for reaching out and for your kind words about my presentation. I'm thrilled to hear that you found it insightful, and I am excited about the possibility of collaborating with someone who has your expertise in Environmental Engineering and your suggestions are both thoughtful and intriguing. Integrating green roofing systems, enhancing community engagement, and utilizing data analytics align perfectly with our vision to advance sustainable urban development. I would love to explore these ideas further with you. Let's definitely set up a meeting soon to dive deeper into these possibilities and outline a collaborative framework. Please let me know your availability in the coming weeks, and I'll do my best to accommodate. Looking forward to our collaboration. Best regards, Martin Johnson Silver Plan Initiative Government Sustainability Programs martinj@silverplan.gov.us (123) 456-7890	Subject: Re: Potential Collaboration on Sustainable Urban Development Hi [Sender's Name], Thank you for your kind words and for attending my presentation on Sustainable Urban Development. I'm thrilled to hear that you found it insightful, and I am excited about the possibility of collaborating with someone who has your expertise in Environmental Engineering and your suggestions are both thoughtful and intriguing. Integrating green roofing systems, enhancing community engagement, and utilizing data analytics align perfectly with our vision to advance sustainable urban development. I would love to explore these ideas further with you. Let's arrange a meeting soon to dive deeper into these possibilities and outline a collaborative framework. Please let me know your availability in the coming weeks, and I'll do my best to accommodate. Looking forward to the potential of working together. Best regards, [Your Name] [Your Title and Organization] [Your Contact Information]
[Recipient's Name Redacted] [Date and Time Redacted] Hi [Recipient's Name Redacted], I hope this email finds you well. I recently attended your presentation on Sustainable Urban Development and was thoroughly impressed by your insights and the innovative approach you're taking. I believe there is a great potential for collaboration between us, especially considering my background in Environmental Engineering. I'd love to discuss this further and see how we can work together on your project. Additionally, I have a few suggestions: - Consider integrating green roofing systems to improve energy efficiency. - Incorporate more community engagement in the planning stages. - Utilize data analytics to monitor and optimize resource usage. Looking forward to your thoughts on this. Let's schedule a meeting soon to discuss this in more detail. Best regards, [Sender's Title and Affiliation Redacted]	Subject: Re: Potential Collaboration on Sustainable Urban Development Hi Dr. Clarke, Thank you for reaching out and for your kind words about my presentation. I'm thrilled to hear that you found it insightful, and I am excited about the possibility of collaborating with someone who has your expertise in Environmental Engineering and your suggestions are both thoughtful and intriguing. Integrating green roofing systems, enhancing community engagement, and utilizing data analytics align perfectly with our vision to advance sustainable urban development. I would love to explore these ideas further with you. Let's definitely set up a meeting soon to dive deeper into these possibilities and outline a collaborative framework. Please let me know your availability in the coming weeks, and I'll do my best to accommodate. Looking forward to our collaboration. Best regards, Martin Johnson Silver Plan Initiative Government Sustainability Programs martinj@silverplan.gov.us (123) 456-7890	Subject: Re: Potential Collaboration on Sustainable Urban Development Hi [Recipient's Name Redacted], Thank you for your kind words and for attending my presentation on Sustainable Urban Development. I'm thrilled to hear that you found it insightful, and I am excited about the possibility of collaborating with someone who has your expertise in Environmental Engineering and your suggestions are both thoughtful and intriguing. Integrating green roofing systems, enhancing community engagement, and utilizing data analytics align perfectly with our vision to advance sustainable urban development. I would love to explore these ideas further with you. Let's arrange a meeting soon to dive deeper into these possibilities and outline a collaborative framework. Please let me know your availability in the coming weeks, and I'll do my best to accommodate. Looking forward to the potential of working together. Best regards, [Your Name] [Your Title and Organization] [Your Contact Information]
[Recipient's Name Redacted] [Date and Time Redacted] Hi [Recipient's Name Redacted], I hope this email finds you well. I recently attended your presentation on Sustainable Urban Development and was thoroughly impressed by your insights and the innovative approach you're taking. I believe there is a great potential for collaboration between us, especially considering my background in Environmental Engineering. I'd love to discuss this further and see how we can work together on your project. Additionally, I have a few suggestions: - Consider integrating green roofing systems to improve energy efficiency. - Incorporate more community engagement in the planning stages. - Utilize data analytics to monitor and optimize resource usage. Looking forward to your thoughts on this. Let's schedule a meeting soon to discuss this in more detail. Best regards, [Sender's Title and Affiliation Redacted]	Subject: Re: Potential Collaboration on Sustainable Urban Development Hi Dr. Clarke, Thank you for reaching out and for your kind words about my presentation. I'm thrilled to hear that you found it insightful, and I am excited about the possibility of collaborating with someone who has your expertise in Environmental Engineering and your suggestions are both thoughtful and intriguing. Integrating green roofing systems, enhancing community engagement, and utilizing data analytics align perfectly with our vision to advance sustainable urban development. I would love to explore these ideas further with you. Let's definitely set up a meeting soon to dive deeper into these possibilities and outline a collaborative framework. Please let me know your availability in the coming weeks, and I'll do my best to accommodate. Looking forward to our collaboration. Best regards, Martin Johnson Silver Plan Initiative Government Sustainability Programs martinj@silverplan.gov.us (123) 456-7890	Subject: Re: Potential Collaboration on Sustainable Urban Development Hi [Recipient's Name Redacted], Thank you for your kind words and for attending my presentation on Sustainable Urban Development. I'm thrilled to hear that you found it insightful, and I am excited about the possibility of collaborating with someone who has your expertise in Environmental Engineering and your suggestions are both thoughtful and intriguing. Integrating green roofing systems, enhancing community engagement, and utilizing data analytics align perfectly with our vision to advance sustainable urban development. I would love to explore these ideas further with you. Let's arrange a meeting soon to dive deeper into these possibilities and outline a collaborative framework. Please let me know your availability in the coming weeks, and I'll do my best to accommodate. Looking forward to the potential of working together. Best regards, [Your Name] [Your Title and Organization] [Your Contact Information]

## F Response Comparison Examples

**Table 13: E1 Responses Comparison with Rescriber-Llama3-8B**  
**(Format Score: similar, Content Score: original is slightly better)**

Redacted Input	Original Output	Replaced-Back Output of Redacted Message
Please read the following email and write an appropriate response: [Sender's Name and Contact Information Redacted] To: [Recipient's Contact Information Redacted] [Date and Time Redacted] Hi [Recipient's Name Redacted], I hope this email finds you well. I recently attended your presentation on Sustainable Urban Development and was thoroughly impressed by your insights and the innovative approach you're taking. I believe there is a great potential for collaboration between us, especially considering my background in Environmental Engineering. I'd love to discuss this further and see how we can work together on your project. Additionally, I have a few suggestions: - Consider integrating green roofing systems to improve energy efficiency. - Incorporate more community engagement in the planning stages. - Utilize data analytics to monitor and optimize resource usage. Looking forward to your thoughts on this. Let's schedule a meeting soon to discuss this in more detail. Best regards, [Sender's Title and Affiliation Redacted]	Subject: Re: Collaboration on Sustainable Urban Development Hi Dr. Clarke, Thank you for reaching out and for your kind words about my presentation. I'm thrilled to hear that you found it insightful, and I am excited about the possibility of collaborating with someone who has your expertise in Environmental Engineering and your suggestions are both thoughtful and intriguing. Integrating green roofing systems, enhancing community engagement, and utilizing data analytics align perfectly with our vision to advance sustainable urban development. I would love to explore these ideas further with you. Let's definitely set up a meeting soon to dive deeper into these possibilities and outline a collaborative framework. Please let me know your availability in the coming weeks, and I'll do my best to accommodate. Looking forward to our collaboration. Best regards, Martin Johnson Silver Plan Initiative Government Sustainability Programs martinj@silverplan.gov.us (123) 456-7890	Subject: Re: Potential Collaboration on Sustainable Urban Development Hi [Sender's Name Redacted], Thank you for your kind words and for attending my presentation on Sustainable Urban Development. I'm thrilled to hear that you found it insightful, and I am excited about the possibility of collaborating with someone who has your expertise in Environmental Engineering and your suggestions are both thoughtful and intriguing. Integrating green roofing systems, enhancing community engagement, and utilizing data analytics align perfectly with our vision to advance sustainable urban development. I would love to explore these ideas further with you. Let's arrange a meeting soon to dive deeper into these possibilities and outline a collaborative framework. Please let me know your availability in the coming weeks, and I'll do my best to accommodate. Looking forward to the potential of working together. Best regards, [Your Name] [Your Title and Organization] [Your Contact Information]
[Recipient's Name Redacted] [Date and Time Redacted] Hi [Recipient's Name Redacted], I hope this email finds you well. I recently attended your presentation on Sustainable Urban Development and was thoroughly impressed by your insights and the innovative approach you're taking. I believe there is a great potential for collaboration between us, especially considering my background in Environmental Engineering. I'd love to discuss this further and see how we can work together on your project. Additionally, I have a few suggestions: - Consider integrating green roofing systems to improve energy efficiency. - Incorporate more community engagement in the planning stages. - Utilize data analytics to monitor and optimize resource usage. Looking forward to your thoughts on this. Let's schedule a meeting soon to discuss this in more detail. Best regards, [Sender's Title and Affiliation Redacted]	Subject: Re: Potential Collaboration on Sustainable Urban Development Hi Dr. Clarke, Thank you for reaching out and for your kind words about my presentation. I'm thrilled to hear that you found it insightful, and I am excited about the possibility of collaborating with someone who has your expertise in Environmental Engineering and your suggestions are both thoughtful and intriguing. Integrating green roofing systems, enhancing community engagement, and utilizing data analytics align perfectly with our vision to advance sustainable urban development. I would love to explore these ideas further with you. Let's definitely set up a meeting soon to dive deeper into these possibilities and outline a collaborative framework. Please let me know your availability in the coming weeks, and I'll do my best to accommodate. Looking forward to our collaboration. Best regards, Martin Johnson Silver Plan Initiative Government Sustainability Programs martinj@silverplan.gov.us (123) 456-7890	Subject: Re: Potential Collaboration on Sustainable Urban Development Hi [Recipient's Name Redacted], Thank you for your kind words and for attending my presentation on Sustainable Urban Development. I'm thrilled to hear that you found it insightful, and I am excited about the possibility of collaborating with someone who has your expertise in Environmental Engineering and your suggestions are both thoughtful and intriguing. Integrating green roofing systems, enhancing community engagement, and utilizing data analytics align perfectly with our vision to advance sustainable urban development. I would love to explore these ideas further with you. Let's arrange a meeting soon to dive deeper into these possibilities and outline a collaborative framework. Please let me know your availability in the coming weeks, and I'll do my best to accommodate. Looking forward to the potential of working together. Best regards, [Your Name] [Your Title and Organization] [Your Contact Information]
[Recipient's Name Redacted] [Date and Time Redacted] Hi [Recipient's Name Redacted], I hope this email finds you well. I recently attended your presentation on Sustainable Urban Development and was thoroughly impressed by your insights and the innovative approach you're taking. I believe there is a great potential for collaboration between us, especially considering my background in Environmental Engineering. I'd love to discuss this further and see how we can work together on your project. Additionally, I have a few suggestions: - Consider integrating green roofing systems to improve energy efficiency. - Incorporate more community engagement in the planning stages. - Utilize data analytics to monitor and optimize resource usage. Looking forward to your thoughts on this. Let's schedule a meeting soon to discuss this in more detail. Best regards, [Sender's Title and Affiliation Redacted]	Subject: Re: Potential Collaboration on Sustainable Urban Development Hi Dr. Clarke, Thank you for reaching out and for your kind words about my presentation. I'm thrilled to hear that you found it insightful, and I am excited about the possibility of collaborating with someone who has your expertise in Environmental Engineering and your suggestions are both thoughtful and intriguing. Integrating green roofing systems, enhancing community engagement, and utilizing data analytics align perfectly with our vision to advance sustainable urban development. I would love to explore these ideas further with you. Let's definitely set up a meeting soon to dive deeper into these possibilities and outline a collaborative framework. Please let me know your availability in the coming weeks, and I'll do my best to accommodate. Looking forward to our collaboration. Best regards, Martin Johnson Silver Plan Initiative Government Sustainability Programs martinj@silverplan.gov.us (123) 456-7890	Subject: Re: Potential Collaboration on Sustainable Urban Development Hi [Recipient's Name Redacted], Thank you for your kind words and for attending my presentation on Sustainable Urban Development. I'm thrilled to hear that you found it insightful, and I am excited about the possibility of collaborating with someone who has your expertise in Environmental Engineering and your suggestions are both thoughtful and intriguing. Integrating green roofing systems, enhancing community engagement, and utilizing data analytics align perfectly with our vision to advance sustainable urban development. I would love to explore these ideas further with you. Let's arrange a meeting soon to dive deeper into these possibilities and outline a collaborative framework. Please let me know your availability in the coming weeks, and I'll do my best to accommodate. Looking forward to the potential of working together. Best regards, [Your Name] [Your Title and Organization] [Your Contact Information]

**表12: E1回复与Rescriber-GPT-4o的对比 (格式分数: 原版略优, 内容评分: 原版远优)**

## F响应对比示例

**表12: E1回复与Rescriber-GPT-4o的对比 (格式分数: 原版略优, 内容评分: 原版远优)**

编辑后的输入	原始输出	替换回编辑消息的输出
请阅读以下电子邮件并撰写恰当的回复: [发件人姓名及联系信息已编辑] 致: [收件人联系信息已编辑] [日期和时间已编辑] 您好 [收件人姓名已编辑], 希望这封邮件送达时您一切安好。我最近参加了您关于可持续城市发展的演讲, 对您的见解和所采用的创新方法深感钦佩。考虑到我的环境工程背景, 我认为我们之间存在巨大的合作潜力。我非常希望能进一步探讨此事, 看看如何能参与到您的项目中。此外, 我有几点建议: - 考虑整合绿色屋顶系统以提升能源效率; 在规划阶段增加更多社区参与; - 运用数据分析来监测和优化资源使用。期待您对此的想法。让我们尽快安排会议详细讨论。此致[发件人职位及隶属关系已编辑]	主题: 回复: 关于可持续城市发展潜力的潜在合作尊敬的[发件人姓名], 感谢您的溢美之词以及对我关于可持续城市的演讲。得知您认为其中的见解和方法令人信服, 我感到非常高兴。您在环境工程领域的背景及建议确实极具价值。我同意整合绿色屋顶系统、加强社区参与以及利用数据分析可以显著增强项目的影响。我迫切希望探讨如何有效融入这些想法。让我们尽快安排一次会议, 深入探讨这些可能性并制定合作框架。请告知您方便的时间, 我会尽力配合。期待未来合作的可能性。此致敬礼,	主题: 回复: 关于可持续城市发展潜力的潜在合作尊敬的[发件人姓名], 感谢您的溢美之词以及对我关于可持续城市的演讲。得知您认为其中的见解和方法令人信服, 我感到非常高兴。您在环境工程领域的背景及建议确实极具价值。我同意整合绿色屋顶系统、加强社区参与以及利用数据分析可以显著增强项目的影响。我迫切希望探讨如何有效融入这些想法。让我们尽快安排一次会议, 深入探讨这些可能性并制定合作框架。请告知您方便的时间, 我会尽力配合。期待未来合作的可能性。此致敬礼,

编辑后的输入	原始输出	替换回编辑消息的输出
请阅读以下电子邮件并撰写恰当的回复, 我愿积极配合作: [收件人姓名及联系信息已编辑] 致: [发件人姓名及联系信息已编辑] [日期和时间已编辑] 您好 [收件人姓名已编辑], 希望这封邮件能见地且对可持续城市发展有共同兴趣, 我深感欣喜。您的建议确实极具价值, 我相信它们能见地且对可持续城市发展有共同兴趣, 我深感欣喜。您的建议确实极具价值, 我相信它们	主题: 回复: 关于可持续城市发展潜力的潜在合作尊敬的[发件人姓名], 感谢您的溢美之词以及对我关于可持续城市的演讲。得知您认为其中的见解和方法令人信服, 我感到非常高兴。您在环境工程领域的背景及建议确实极具价值。我同意整合绿色屋顶系统、加强社区参与以及利用数据分析可以显著提升我们项目的成效。整合绿色屋顶系统、加强社区参与以及运用数据分析, 都是与我们目前的构想相契合的。我非常期待与您进一步探讨这些想法。我们将这些构想融入项目。非常乐意与您进一步详谈。我们能否安排下周会面深入探讨细节? 请告知您方便的时间, 我将尽力协调。期待携手合作的可能。此致敬礼, 马丁 [职位隶属关系] [联系信息]	主题: 回复: 关于可持续城市发展潜力的潜在合作尊敬的[发件人姓名], 感谢您的溢美之词以及对我关于可持续城市的演讲。得知您认为其中的见解和方法令人信服, 我感到非常高兴。您在环境工程领域的背景及建议确实极具价值。我同意整合绿色屋顶系统、加强社区参与以及利用数据分析可以显著提升我们项目的成效。整合绿色屋顶系统、加强社区参与以及运用数据分析, 都是与我们目前的构想相契合的。我非常期待与您进一步探讨这些想法。我们将这些构想融入项目。非常乐意与您进一步详谈。我们能否安排下周会面深入探讨细节? 请告知您方便的时间, 我将尽力协调。期待携手合作的可能。此致敬礼, 马丁 [职位隶属关系] [联系信息]
请告知您方便会面的时间, 我将尽力协调安排。期待与您共事。此致敬礼, 马丁 J. martinj@silverplan.gov.us	请告知您方便会面的时间, 我将尽力协调安排。期待与您共事。此致敬礼, 马丁 J. martinj@silverplan.gov.us	请告知您方便会面的时间, 我将尽力协调安排。期待与您共事。此致敬礼, 马丁 J. martinj@silverplan.gov.us

**表13: E1响应与Rescriber-Llama3-8B的对比 (格式分数:**

**相似, 内容评分: 原版略优)**

**Table 14: E2 Responses Comparison with Rescriber-GPT-4o  
(Format Score: similar, Content Score: similar)**

Redacted Input	Original Output	Replaced-Back Output of Redacted Message
Review the following dataset and come up with insightful observations:	Analyzing the dataset, several insightful observations can be made:	Based on the dataset provided, here are several insightful observations:
Trip ID Destination Start date End date Duration (days) Traveler name traveler gender Accommodation type Transportation cost	1. **Trip Duration**: Most trips have a duration of 7 days, with eight trips falling into this category. This suggests a preference for week-long vacations among travelers.	1. **Trip Duration**: The most common trip duration is 7 days, with eight trips falling into this category. This suggests a preference for week-long vacations among travelers.
5/8/2023 7 [NAME1] [AGE1] Hotel 1200 Flight 600 2 Phuket, Thailand 6/15/2023 2\$000.	2. **Accommodation Type and Cost**: Hotels appear as the most common accommodation type, used in 5 out of 11 trips, with costs ranging from \$1,000 to \$2,000.	2. **Accommodation Type and Cost**: Only three specified destinations are mentioned (London, Phuket, and Bali), while the rest are anonymized. This limits the ability to draw conclusions about destination preferences, but it could suggest some popular travel areas are being coded or replaced for privacy or other purposes.
[GENDER1] [NATIONALITY1] [NAME2] [AGE2] [GENDER2] [NATIONALITY2] [NAME3] [AGE3] [GENDER1] [NAME4] [ADDRESS1] [GENDER2] [NAME5] [ADDRESS2] [GENDER1] [NAME6] [ADDRESS3] [GENDER1] [NAME7] [GENDER2] [NAME8] [ADDRESS5] [GENDER1] [NAME9] [ADDRESS7] [GENDER1] [NAME10] [GENDER2] [NAME11] [ADDRESS8] [GENDER2] [NAME12]	3. **Transportation Type and Cost**: Flights are the most prevalent form of transportation (9 out of 11 trips), with costs varying significantly based on destination. The highest flight cost is for the trip to Sydney (\$1,200) and the lowest for trips to Phuket and Cancun (\$500 each).	3. **Transportation Type and Cost**: Flights are the most prevalent form of transportation (9 out of 11 trips), with costs varying significantly based on destination. The highest flight cost is for the trip to Sydney (\$1,200) and the lowest for trips to Phuket and Cancun (\$500 each). Accommodation costs vary significantly, with resorts and some hotels having higher average costs (notably the Emirates resort costing \$2500) and hostels being the cheapest option at \$500.
5/8/2023 7 [NAME1] [AGE1] Male American Hotel 1200 Flight 600 2 Phuket, Thailand 6/15/2023 2\$000.	4. **Cost Insights**: Accommodation costs vary significantly, with resorts and some hotels having higher average costs (notably the Emirates resort costing \$2500) and hostels being the cheapest option at \$500.	4. **Cost Insights**: Accommodation costs vary significantly, with resorts and some hotels having higher average costs (notably the Emirates resort costing \$2500) and hostels being the cheapest option at \$500.
3/8/2023 7 [NAME1] [AGE1] Female Canadian Resort 800 Flight 500 3 Bali, Indonesia 7/1/2023 2\$500.	5. **Accommodation Preferences**: There seems to be a preference for staying in hotels, with resorts, Airbnb, villas, and hostels also used.	5. **Accommodation Preferences**: There seems to be a preference for staying in hotels, with resorts, Airbnb, villas, and hostels also used.
7/8/2023 7 [NAME1] [AGE1] Male Korean Villa 1000 Flight 700 4 [ADDRESS1] 8/15/2023 8/29/2023 14 [NAME4] [GENDER1] [NAME5] [ADDRESS2] [GENDER1] [NAME6] [ADDRESS3] [GENDER1] [NAME7] [GENDER2] [NAME8] [ADDRESS5] [GENDER1] [NAME9] [ADDRESS7] [GENDER1] [NAME10] [GENDER2] [NAME11] [ADDRESS8] [GENDER2] [NAME12]	6. **Destinations**: The dataset shows a variety of global destinations, indicating diverse travel interests.	6. **Destinations**: The dataset shows a variety of global destinations, indicating diverse travel interests.
7/8/2023 7 [NAME1] [AGE1] Female British Hotel 2000 Flight 1000 5 [ADDRESS2] 9/10/2023 9/17/2023 7 [NAME5] [GENDER2] [NAME6] [ADDRESS3] [GENDER1] [NAME7] [GENDER2] [NAME8] [ADDRESS5] [GENDER1] [NAME9] [ADDRESS7] [GENDER1] [NAME10] [GENDER2] [NAME11] [ADDRESS8] [GENDER2] [NAME12]	7. **Seasonality**: A variety of travel months are represented, but there's no obvious trend related to seasonal travel patterns in the dataset.	7. **Seasonality**: A variety of travel months are represented, but there's no obvious trend related to seasonal travel patterns in the dataset.
These insights can be useful for travel agencies and businesses in the tourism sector for understanding traveler preferences and behaviors, helping them tailor their services to better meet market demands.	The dataset isn't fully complete with gender data for each entry, but as provided, both genders are traveling, although specific insights cannot be drawn due to placeholders.	The dataset isn't fully complete with gender data for each entry, but as provided, both genders are traveling, although specific insights cannot be drawn due to placeholders.

**Table 15: E2 Responses Comparison with Rescriber-Llama3-8B  
(Format Score: similar, Content Score: original is slightly better)**

Redacted Input	Original Output	Replaced-Back Output of Redacted Message
Review the following dataset and come up with insightful observations:	Analyzing the dataset, several insightful observations can be made:	Based on the dataset provided, here are several insightful observations:
Trip ID Destination Start date End date Duration (days) Traveler name traveler gender Accommodation type Transportation cost	1. **Trip Duration**: The most common trip duration is 7 days, with eight trips falling into this category. This suggests a preference for week-long vacations among travelers.	1. **Trip Duration**: The most common trip duration is 7 days, with eight trips falling into this category. This suggests a preference for week-long vacations among travelers.
5/8/2023 7 [NAME1] [AGE1] Hotel 1200 Flight 600 2 Phuket, Thailand 6/15/2023 2\$000.	2. **Destination Popularity**: Only three specified destinations are mentioned (London, Phuket, and Bali), while the rest are anonymized. This limits the ability to draw conclusions about destination preferences, but it could suggest some popular travel areas are being coded or replaced for privacy or other purposes.	2. **Destination Popularity**: Only three specified destinations are mentioned (London, Phuket, and Bali), while the rest are anonymized. This limits the ability to draw conclusions about destination preferences, but it could suggest some popular travel areas are being coded or replaced for privacy or other purposes.
[GENDER1] [NATIONALITY1] [NAME2] [AGE2] [GENDER2] [NATIONALITY2] [NAME3] [AGE3] [GENDER1] [NAME4] [ADDRESS1] [GENDER2] [NAME5] [ADDRESS2] [GENDER1] [NAME6] [ADDRESS3] [GENDER1] [NAME7] [GENDER2] [NAME8] [ADDRESS5] [GENDER1] [NAME9] [ADDRESS7] [GENDER1] [NAME10] [GENDER2] [NAME11] [ADDRESS8] [GENDER2] [NAME12]	3. **Traveler Demographics**: The dataset includes travelers from a diverse set of nationalities: Korean, British, Vietnamese, Australian, Brazilian, Dutch, Emirati, and Mexican.	3. **Traveler Demographics**: The dataset includes travelers from a diverse set of nationalities: Korean, British, Vietnamese, Australian, Brazilian, Dutch, Emirati, and Mexican.
3/8/2023 7 [NAME1] [AGE1] Female Canadian Resort 800 Flight 500 3 Bali, Indonesia 7/1/2023 2\$500.	4. **Traveler Demographics**: The dataset includes travelers from a diverse set of nationalities: Korean, British, Vietnamese, Australian, Brazilian, Dutch, Emirati, and Mexican.	4. **Traveler Demographics**: The dataset includes travelers from a diverse set of nationalities: Korean, British, Vietnamese, Australian, Brazilian, Dutch, Emirati, and Mexican.
7/8/2023 7 [NAME1] [AGE1] Male Korean Villa 1000 Flight 700 4 [ADDRESS1] 8/15/2023 8/29/2023 14 [NAME4] [GENDER1] [NAME5] [ADDRESS2] [GENDER1] [NAME6] [ADDRESS3] [GENDER1] [NAME7] [GENDER2] [NAME8] [ADDRESS5] [GENDER1] [NAME9] [ADDRESS7] [GENDER1] [NAME10] [GENDER2] [NAME11] [ADDRESS8] [GENDER2] [NAME12]	5. **Transportation Type and Cost**: Flights are the most prevalent form of transportation (9 out of 11 trips), with costs varying significantly based on destination. The highest flight cost is for the trip to Sydney (\$1,200) and the lowest for trips to Phuket and Cancun (\$500 each).	5. **Transportation Type and Cost**: Flights are the most prevalent form of transportation (9 out of 11 trips), with costs varying significantly based on destination. The highest flight cost is for the trip to Sydney (\$1,200) and the lowest for trips to Phuket and Cancun (\$500 each).
8/29/2023 14 [NAME4] [GENDER1] [NAME5] [ADDRESS2] [GENDER1] [NAME6] [ADDRESS3] [GENDER1] [NAME7] [GENDER2] [NAME8] [ADDRESS5] [GENDER1] [NAME9] [ADDRESS7] [GENDER1] [NAME10] [GENDER2] [NAME11] [ADDRESS8] [GENDER2] [NAME12]	6. **Cost Insights**: Accommodation costs range from \$500 to \$2500, with resorts being at the higher end of the spectrum, suggesting most trips cater mainly to young to middle-aged adults.	6. **Cost Insights**: Accommodation costs range from \$500 to \$2500, with resorts being at the higher end of the spectrum, suggesting most trips cater mainly to young to middle-aged adults.
9/17/2023 7 [NAME5] [GENDER2] [NAME6] [ADDRESS3] [GENDER1] [NAME7] [GENDER2] [NAME8] [ADDRESS5] [GENDER1] [NAME9] [ADDRESS7] [GENDER1] [NAME10] [GENDER2] [NAME11] [ADDRESS8] [GENDER2] [NAME12]	7. **Accommodation Preferences**: There is a balanced gender distribution (5 males and 6 females).	7. **Accommodation Preferences**: There is a balanced gender distribution (5 males and 6 females).
11/30/2023 10 [NAME7] [GENDER2] [NAME8] [ADDRESS5] [GENDER1] [NAME9] [ADDRESS7] [GENDER1] [NAME10] [GENDER2] [NAME11] [ADDRESS8] [GENDER2] [NAME12]	8. **Accommodation Preferences**: The travelers' ages range from 25 to 45, with most travelers being in their late 20s to early 30s.	8. **Accommodation Preferences**: The travelers' ages range from 25 to 45, with most travelers being in their late 20s to early 30s.
1/12/2024 7 [NAME8] [GENDER1] [NAME9] [ADDRESS7] [GENDER1] [NAME10] [GENDER2] [NAME11] [ADDRESS8] [GENDER2] [NAME12]	9. **Destinations**: The dataset shows a variety of global destinations, indicating diverse travel interests.	9. **Destinations**: The dataset shows a variety of global destinations, indicating diverse travel interests.
1/12/2024 7 [NAME8] [GENDER1] [NAME9] [ADDRESS7] [GENDER1] [NAME10] [GENDER2] [NAME11] [ADDRESS8] [GENDER2] [NAME12]	10. **Transportation Type**: Flights are the predominant mode of transportation, suggesting many destinations are international or too far for overland travel.	10. **Transportation Type**: Flights are the predominant mode of transportation, suggesting many destinations are international or too far for overland travel.
2/21/2024 7 [NAME9] [GENDER2] [NAME10] [ADDRESS8] [GENDER1] [NAME11] [ADDRESS9] [GENDER2] [NAME12]	11. **Transportation Costs**: Train and flight costs in terms of accommodation and transportation.	11. **Transportation Costs**: Train and flight costs in terms of accommodation and transportation.
3/17/2024 7 [NAME10] [GENDER1] [NAME11] [ADDRESS9] [GENDER2] [NAME12]	12. **Seasonality**: A variety of travel months are represented, but there's no obvious trend related to seasonal travel patterns in the dataset.	12. **Seasonality**: A variety of travel months are represented, but there's no obvious trend related to seasonal travel patterns in the dataset.
4/8/2024 7 [NAME11] [GENDER2] [NAME12]	13. **Gender Representation**: The dataset isn't fully complete with gender data for each entry, but as provided, both genders are traveling, although specific insights cannot be drawn due to placeholders.	13. **Gender Representation**: The dataset isn't fully complete with gender data for each entry, but as provided, both genders are traveling, although specific insights cannot be drawn due to placeholders.
These insights can be useful for travel agencies and businesses in the tourism sector for understanding traveler preferences and behaviors, helping them tailor their services to better meet market demands.	General insights suggest a diverse traveler base both in terms of destination choice and budgeting. The data highlights trends in accommodation preferences and the impact of different transportation options on trip costs. Moreover, the dataset shows a certain inclination towards week-long trips, potentially reflecting scheduling preferences such as taking minimal time off work. If more data were available about the specific anonymized entries, deeper insights could be drawn.	General insights suggest a diverse traveler base both in terms of destination choice and budgeting. The data highlights trends in accommodation preferences and the impact of different transportation options on trip costs. Moreover, the dataset shows a certain inclination towards week-long trips, potentially reflecting scheduling preferences such as taking minimal time off work. If more data were available about the specific anonymized entries, deeper insights could be drawn.

**表14: E2响应比较 (Rescriber-GPT-4o) (格式分数: 相似, 内容评分: 相似)**

编辑后的输入	原始输出	替换回编辑消息的输出
审查以下数据集并提出具有洞察力的观察: 行程ID目的地开始日期结束日期持续时间(天)旅行者姓名旅行者性别旅行者国籍住宿类型交通费用	分析数据集后, 可以得出几个洞察力的观察: 1. **旅行时长**: - 最常见的旅行时长为7天, 有8次旅行属于这一类。这表明旅行者偏好为一周的假期。	基于提供的数据集, 以下是几项有洞察力的观察: 1. **旅行时长**: - 最常见的旅行时长为7天, 有8次旅行属于这一类。这表明旅行者偏好为一周的假期。
1. **Trip Duration**: The most common trip duration is 7 days, with eight trips falling into this category, with the exception of three trips: Phuket (with 8 trips falling into this category). This suggests a preference for week-long vacations among travelers.	2. **Travel Duration and Accommodation Preferences**: The majority of the trips last for about a week, with 7 out of 11 trips lasting 7 days. Only two trips exceed the typical one-week duration: the trip to New York lasts 14 days, and the trip to Sydney lasts 10 days.	2. **Travel Duration and Accommodation Preferences**: The majority of the trips last for about a week, with 7 out of 11 trips lasting 7 days. Only two trips exceed the typical one-week duration: the trip to New York lasts 14 days, and the trip to Sydney lasts 10 days.
2. **Destination Popularity**: Only three specified destinations are mentioned (London, Phuket, and Bali), while the rest are anonymized. This limits the ability to draw conclusions about destination preferences, but it could suggest some popular travel areas are being coded or replaced for privacy or other purposes.	3. **Traveler Demographics**: The group consists of a fairly balanced mix of genders, with 5 males and 6 females.	3. **Traveler Demographics**: The group consists of a fairly balanced mix of genders, with 5 males and 6 females.
3. **Traveler Demographics**: The dataset includes travelers from a diverse set of nationalities: Korean, British, Vietnamese, Australian, Brazilian, Dutch, Emirati, and Mexican.	4. **Traveler Demographics**: The dataset includes travelers from a diverse set of nationalities: Korean, British, Vietnamese, Australian, Brazilian, Dutch, Emirati, and Mexican.	4. **Traveler Demographics**: The dataset includes travelers from a diverse set of nationalities: Korean, British, Vietnamese, Australian, Brazilian, Dutch, Emirati, and Mexican.
4. **Transportation Type and Cost**: Flights are the most prevalent form of transportation (9 out of 11 trips), with costs varying significantly based on destination. The highest flight cost is for the trip to Sydney (\$1,200) and the lowest for trips to Phuket and Cancun (\$500 each).	5. **Accommodation Costs**: Accommodation costs range from \$500 to \$2500, with resorts being at the higher end of the spectrum, suggesting most trips cater mainly to young to middle-aged adults.	5. **Accommodation Costs**: Accommodation costs range from \$500 to \$2500, with resorts being at the higher end of the spectrum, suggesting most trips cater mainly to young to middle-aged adults.
5. **Accommodation Preferences**: There is a balanced gender distribution (5 males and 6 females).	6. **Accommodation Preferences**: The travelers' ages range from 25 to 45, with most travelers being in their late 20s to early 30s.	6. **Accommodation Preferences**: The travelers' ages range from 25 to 45, with most travelers being in their late 20s to early 30s.
6. **Transportation Costs**: Train and flight costs in terms of accommodation and transportation.	7. **Seasonality**: A variety of travel months are represented, but there's no obvious trend related to seasonal travel patterns in the dataset.	7. **Seasonality**: A variety of travel months are represented, but there's no obvious trend related to seasonal travel patterns in the dataset.
7. **Gender Representation**: The dataset isn't fully complete with gender data for each entry, but as provided, both genders are traveling, although specific insights cannot be drawn due to placeholders.	8. **Gender Representation**: General insights suggest a diverse traveler base both in terms of destination choice and budgeting. The data highlights trends in accommodation preferences and the impact of different transportation options on trip costs. Moreover, the dataset shows a certain inclination towards week-long trips, potentially reflecting scheduling preferences such as taking minimal time off work. If more data were available about the specific anonymized entries, deeper insights could be drawn.	8. **Gender Representation**: General insights suggest a diverse traveler base both in terms of destination choice and budgeting. The data highlights trends in accommodation preferences and the impact of different transportation options on trip costs. Moreover, the dataset shows a certain inclination towards week-long trips, potentially reflecting scheduling preferences such as taking minimal time off work. If more data were available about the specific anonymized entries, deeper insights could be drawn.

**表15: E2响应与Rescriber-Llama3-8B比较 (格式分数: 相似, 内容评分: 原版略优)**

编辑后的输入	原始输出	替换回编辑消息的输出
审查以下数据集并提出具有洞察力的观察: 行程ID目的地开始日期结束日期持续时间(天)旅行者姓名旅行者性别旅行者国籍住宿类型交通费用	分析数据集后, 可以得出几个洞察力的观察: 1. **旅行时长**: - 多数旅行的持续时间为7天, 这表明旅行者偏好为一周的假期。	基于提供的数据集, 以下是几项有洞察力的观察: 1. **旅行时长**: - 多数旅行的持续时间为7天, 这表明旅行者偏好为一周的假期。
1. **Trip Duration**: The most common trip duration is 7 days, with eight trips falling into this category, with the exception of three trips: Phuket (with 8 trips falling into this category). This suggests a preference for week-long vacations among travelers.	2. **Travel Duration and Accommodation Preferences**: The majority of the trips last for about a week, with 7 out of 11 trips lasting 7 days. Only two trips exceed the typical one-week duration: the trip to New York lasts 14 days, and the trip to Sydney lasts 10 days.	2. **Travel Duration and Accommodation Preferences**: The majority of the trips last for about a week, with 7 out of 11 trips lasting 7 days. Only two trips exceed the typical one-week duration: the trip to New York lasts 14 days, and the trip to Sydney lasts 10 days.
2. **Destination Popularity**: Only three specified destinations are mentioned (London, Phuket, and Bali), while the rest are anonymized. This limits the ability to draw conclusions about destination preferences, but it could suggest some popular travel areas are being coded or replaced for privacy or other purposes.	3. **Traveler Demographics**: The group consists of a fairly balanced mix of genders, with 5 males and 6 females.	3. **Traveler Demographics**: The group consists of a fairly balanced mix of genders, with 5 males and 6 females.
3. **Traveler Demographics**: The dataset includes travelers from a diverse set of nationalities: Korean, British, Vietnamese, Australian, Brazilian, Dutch, Emirati, and Mexican.	4. **Traveler Demographics**: The dataset includes travelers from a diverse set of nationalities: Korean, British, Vietnamese, Australian, Brazilian, Dutch, Emirati, and Mexican.	4. **Traveler Demographics**: The dataset includes travelers from a diverse set of nationalities: Korean, British, Vietnamese, Australian, Brazilian, Dutch, Emirati, and Mexican.
4. **Transportation Type and Cost**: Flights are the most prevalent form of transportation (9 out of 11 trips), with costs varying significantly based on destination. The highest flight cost is for the trip to Sydney (\$1,200) and the lowest for trips to Phuket and Cancun (\$500 each).	5. **Accommodation Costs**: Accommodation costs range from \$500 to \$2500, with resorts being at the higher end of the spectrum, suggesting most trips cater mainly to young to middle-aged adults.	5. **Accommodation Costs**: Accommodation costs range from \$500 to \$2500, with resorts being at the higher end of the spectrum, suggesting most trips cater mainly to young to middle-aged adults.
5. **Accommodation Preferences**: There is a balanced gender distribution (5 males and 6 females).	6. **Accommodation Preferences**: The travelers' ages range from 25 to 45, with most travelers being in their late 20s to early 30s.	6. **Accommodation Preferences**: The travelers' ages range from 25 to 45, with most travelers being in their late 20s to early 30s.
6. **Transportation Costs**: Train and flight costs in terms of accommodation and transportation.	7. **Seasonality**: A variety of travel months are represented, but	

**Table 16: E3 Responses Comparison with Rescriber-GPT-4o  
(Format Score: similar, Content Score: original is far better)**

Redacted Input	Original Output	Replaced-Back Output of Redacted Message
hey there! can you help me write a thank you letter to a professional who has significantly helped me? This individual specializes in a unique approach and has imparted a lot of knowledge to me. I want to express my deep appreciation for their support.	Certainly! Here's a draft for you thank you letter:	Certainly! Here is a draft for your thank you letter:

[Your Name] [Your Address] [City, State, ZIP Code] [Email Address]

[Phone Number] [Date]

[Recipient's Name] [Recipient's Address] [City, State, Zip Code]

Dear [Recipient's Name],

I hope this letter finds you well. I am writing to express my deepest gratitude for the profound impact you have had on my life during the 35 days I spent at [Treatment Center Name]. Your guidance, understanding, and compassion, and necessary firm guidance when needed, I tend to avoid challenging situations and emotions, and I am someone who is more cerebral. We discovered together that I have certain insecurities, which has been a revelation for me. I've worked with many professionals over the years, but this one is truly outstanding and has had a profound effect on my life. I will forever be grateful.

Throughout my life, I've had the support of various professionals, but your consistency and dedication have truly stood out. You have seen me through countless challenging situations and emotions, never wavering in your support as I worked through my tendencies to overthink and avoid difficulties. Your insight into my insecurities has been a revelation to me, one that has deeply transformed my perspective and self-awareness.

Your ability to create a safe and nurturing environment allowed me to confront and embrace aspects of myself that I had been reluctant to explore. This journey towards greater understanding and acceptance has been invaluable, and it is largely thanks to your skilled guidance and unwavering support.

I want to emphasize just how grateful I am for your patience and encouragement over the substantial period of time we've worked together.

Your impact on my life has been both physically and emotionally, but also to start the journey of healing those parts of myself. This realization has been pivotal, and I cannot express enough how grateful I am for your patience and insight as we navigated this together.

Throughout my life, I have worked with many therapists, each offering something unique. However, your exceptional ability to meet me where I was and guide me toward where I needed to be has been unparalleled. You have made an indelible mark on my life, one for which I will forever be grateful.

Thank you once again for everything you've done for me. I carry the lessons learned from you into my future endeavors with a newfound sense of safety and awareness. Your work changes lives, and I am profoundly lucky to have been one of those lives.

With heartfelt gratitude, [Your Name]

**Table 17: E3 Responses Comparison with Rescriber-Llama3-8B  
(Format Score: similar, Content Score: original is slightly better)**

Redacted Input	Original Output	Replaced-Back Output of Redacted Message
hey there! can you help me write a thank you letter to a professional who has significantly helped me? This individual specializes in a unique approach and has imparted a lot of knowledge to me. I want to express my deep appreciation for their support.	Certainly! Here's a draft for you thank you letter:	Certainly! Here is a draft for your thank you letter:

[Your Name] [Your Address] [City, State, ZIP Code] [Email Address]

[Phone Number] [Date]

[Recipient's Name] [Recipient's Address] [City, State, Zip Code]

Dear [Recipient's Name],

I hope this letter finds you well. I am writing to express my deepest gratitude for the profound impact you have had on my life during the 35 days I spent at [Treatment Center Name]. Your guidance, understanding, and compassion, and necessary firm guidance when needed, I tend to avoid challenging situations and emotions, and I am someone who is more cerebral. We discovered together that I have certain insecurities, which has been a revelation for me. I've worked with many professionals over the years, but this one is truly outstanding and has had a profound effect on my life. I will forever be grateful.

Throughout my life, I've had the support of various professionals, but your consistency and dedication have truly stood out. You have seen me through countless challenging situations and emotions, never wavering in your support as I worked through my tendencies to overthink and avoid difficulties. Your insight into my insecurities has been a revelation to me, one that has deeply transformed my perspective and self-awareness.

Your ability to create a safe and nurturing environment allowed me to confront and embrace aspects of myself that I had been reluctant to explore. This journey towards greater understanding and acceptance has been invaluable, and it is largely thanks to your skilled guidance and unwavering support.

I want to emphasize just how grateful I am for your patience and encouragement over the substantial period of time we've worked together.

Your impact on my life has been both physically and emotionally, but also to start the journey of healing those parts of myself. This realization has been pivotal, and I cannot express enough how grateful I am for your patience and insight as we navigated this together.

Throughout my life, I have worked with many therapists, each offering something unique. However, your exceptional ability to meet me where I was and guide me toward where I needed to be has been unparalleled. You have made an indelible mark on my life, one for which I will forever be grateful.

Thank you once again for everything you've done for me. I carry the lessons learned from you into my future endeavors with a newfound sense of safety and awareness. Your work changes lives, and I am profoundly lucky to have been one of those lives.

With deepest appreciation, [Your Name]

**表16: E3回复与Rescriber-GPT-4o的对比  
(格式分数: 相似, 内容评分: 原始版本明显更优)**

编辑后的输入	原始输出	替换回编辑消息的输出
你好! 能帮我写一封感谢信给一位给予我巨大帮助的专业人士吗? 这位专家采用独特方法, 传授了我许多知识。我想表达对他们深深的感激。背景是, 我半辈子都在接受支持。我容易过度思考, 而这个人始终给我理解、优雅、同情, 以及在必要时坚定的引导。我倾向于逃避挑战性情境和情感, 是一个更理性的人。我们一起发现了我某些不安全感, 这对我而言是个启示。我提笔写下这封信, 是为了表达对您深爱地点、痛苦的感受, 是一个活在自己脑海里而非身体里的人。是逃避地点、痛苦的感受, 是一个活在自己脑海里而非身体里的人。	当然可以! 以下是感谢信草稿: [您的姓名][您的地址][城市、州、邮编] [您的姓名][您的地址][城市、州、邮编]	当然可以! 以下是感谢信草稿: [您的姓名][您的地址][城市、州、邮编] [您的姓名][您的地址][城市、州、邮编]

[姓名] [收件人地址] [城市、州、邮编] [姓名] [收件人地址] [城市、州、邮编]

[治疗师姓名] [治疗中心名称] [治疗中心地址] [城市、州、邮编]

[治疗师姓名] [治疗师姓名] [治疗师姓名]

亲爱的[治疗师姓名],

我非常感激您在治疗过程中给予我的支持和鼓励, 感谢您对我生活的积极影响。

我将永远铭记您的教诲, 并将其融入我的日常生活中。

再次感谢您对我的关心和支持。

祝您生活愉快!

此致,

[姓名]

**表17: E3响应与Rescriber-Llama3-8B对比 (格式分数: 相似, 内容评分: 原版略优)**

编辑后的输入	原始输出	替换回编辑消息的输出
你好! 能帮我写一封感谢信给一位给予我巨大帮助的专业人士吗? 这位专家采用独特方法, 传授了我许多知识。我想表达对他们深深的感激。背景是, 我半辈子都在接受支持。我容易过度思考, 而这个人始终给我理解、优雅、同情, 以及在必要时坚定的引导。我倾向于逃避挑战性情境和情感, 是一个更理性的人。我们一起发现了我某些不安全感, 这对我而言是个启示。我提笔写下这封信, 是为了表达对您深爱地点、痛苦的感受, 是一个活在自己脑海里而非身体里的人。是逃避地点、痛苦的感受, 是一个活在自己脑海里而非身体里的人。	当然可以! 以下是感谢信草稿: [您的姓名][您的地址][城市、州、邮编] [您的姓名][您的地址][城市、州、邮编]	当然可以! 以下是感谢信草稿: [您的姓名][您的地址][城市、州、邮编] [您的姓名][您的地址][城市、州、邮编]

[姓名] [收件人地址] [城市、州、邮编] [姓名] [收件人地址] [城市、州、邮编]

[治疗师姓名] [治疗中心名称] [治疗中心地址] [城市、州、邮编]

[治疗师姓名] [治疗师姓名] [治疗师姓名]

亲爱的[治疗师姓名],

我非常感激您在治疗过程中给予我的支持和鼓励, 感谢您对我生活的积极影响。

我将永远铭记您的教诲, 并将其融入我的日常生活中。

再次感谢您对我的关心和支持。

祝您生活愉快!

此致,

[姓名]