From Patient Burdens to User Agency: Designing for Real-Time Protection Support in Online Health Consultations

Shuning Zhang
zsn23@mails.tsinghua.edu.cn
Tsinghua University
Beijing, China

Ying Ma
School of Computing and Information Systems
University of Melbourne
Melbourne, Australia
ying.ma1@student.unimelb.edu.au

Yongquan 'Owen' Hu
yongquan@ahlab.org
Augmented Human Lab
National University of Singapore
Singapore

Ting Dang
ting.dang@unimelb.edu.au
University of Melbourne
Melbourne, VIC, Australia

Hong Jia
hong.jia@auckland.ac.nz
University of Auckland
Auckland, New Zealand

Xin Yi*
yixin@tsinghua.edu.cn
Tsinghua University
Beijing, China

Hewu Li
lihewu@cernet.edu.cn
Tsinghua University
Beijing, China

Abstract

Online medical consultation platforms, while convenient, are undermined by significant privacy risks that erode user trust. We first conducted in-depth semi-structured interviews with 12 users to understand their perceptions of security and privacy landscapes on online medical consultation platforms, as well as their practices, challenges and expectation. Our analysis reveals a critical disconnect between users' desires for anonymity and control, and platform realities that offload the responsibility of "privacy labor". To bridge this gap, we present SafeShare, an interaction technique that leverages localized LLM to redact consultations in real-time. SafeShare balances utility and privacy through selectively anonymize private information. A technical evaluation of SafeShare's core PII detection module on 3 dataset demonstrates high efficacy, achieving 89.64% accuracy with Qwen3-4B on IMCS21 dataset.

CCS Concepts

• **Human-centered computing** → *Human computer interaction (HCI)*; • **Applied computing** → **Health informatics**; • **Security and privacy** → **Privacy protections**.

Keywords

Online consultation, Medical consultation, Health, Privacy

*Corresponding author.

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference Format:
Shuning Zhang, Ying Ma, Yongquan 'Owen' Hu, Ting Dang, Hong Jia, Xin Yi, and Hewu Li. 2025. From Patient Burdens to User Agency: Designing for Real-Time Protection Support in Online Health Consultations. In . ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Online medical consultation [2], a service allowing patients to seek remote advice from healthcare professionals, has seen explosive growth, especially during and after the COVID-19 pandemic. The online consultation market reached RMB 63.62 billion by 2024 globally [5]. This prevalent form, typically involving communication via text, audio or video, allows patients to describe symptoms and receive diagnoses or treatment plans from qualified practitioners without a physical visit.

However, this convenience also comes with severe privacy concerns. Various news have unveiled that online consultation apps has privacy leakage issues [18, 25], and patients are concerned about their own privacy leakage [32, 42]. Worse still, some specific patients already experienced the harm caused by the privacy and security harm of online medical consultation [34, 38]. While these problems are known, a deep, user-centered understanding of how patients perceive these risks and attempt to manage their privacy is underexplored, yet critical for building the trust necessary for sustainable adoption. Therefore, our investigation is guided by the following research questions: (RQ1) What are the perceived privacy risks associated with online medical consultations? (RQ2) What are users' privacy protection practices and expectations? (RQ3) How can a real-time, localized anonymization technique effectively mitigate the aforementioned privacy risks?

To answer these questions, we first conducted an in-depth, semi-structured interview (N=12). We then designed SafeShare, and evaluated its technical implementation using three datasets. For RQ1, we found user apprehensions are not confined to immediate data leakage but extend to the creation of permanent, algorithmic health

profiles, which include three primary risk perceptions: (1) platform-level data exploitation, (2) unauthorized disclosure by individual practitioners, and (3) general distrust stemming from profit motives. These concerns are framed by a "tripartite model of trust", where users tend to trust the professionalism of individual doctors but remain skepticism towards the platform operators.

For RQ2, we found that users have significant "privacy labor" to manage their data, a responsibility offloaded onto them by platforms, including proactive self-censorship such as using pseudonyms, manually cropping identifying details from photos, and blurring information on medical reports. However, they perceive these efforts as both burdensome and ultimately insufficient, leading to "privacy fatigue". Therefore, users have clear expectations for system-level reforms, demanding (1) granular control over their data, including the right to be forgotten, (2) enhanced in-context transparency and accountability mechanisms, and (3) robust external regulation, as they do not trust platforms to be self-disciplined.

Finally, for RQ3, we introduce SafeShare, an interaction technique that acts as an intelligent agent within the consultation interface. SafeShare leverages a localized LLM to automate the redaction of Personal Identifiable Information (PII) and provide real-time, context-aware justifications for doctors' data requests, thereby reducing user burden and enhancing transparency. Our technical evaluation shows that the core anonymization module is highly effective, achieving an accuracy of 89.64% with Qwen3-4B on IMCS21 [6] dataset in identifying sensitive information in clinical text. Therefore, this paper makes two primary contributions:

- We provide the first formal and in-depth qualitative characterization of users' perceived privacy risks, mitigating strategies, and expectations in online health consultations.
- We design and evaluate SafeShare, an automatic anonymization technique in medical context capable of running on local devices.

2 Related Work

Our work is situated within the broader research landscape of user trust in online healthcare, the privacy dynamics of medical consultation, and the design of privacy-enhancing technologies. A primary challenge in online health services is the establishment of user trust, which is often more fragile in digital settings compared to face-to-face interactions [30]. Trust is a critical mediator influencing a user's willingness to disclose sensitive health information. For example, users show greater willingness to share data with trusted human doctors than with AI systems [33]. This dynamic frequently forces users into a "privacy calculus" [17], where they must weigh the diagnostic benefits of sharing data against the perceived privacy risks [39]. Our paper substantiates this, revealing a clear pattern where users willingly share diagnostic information but strongly resist disclosing personal identifiable information (PII). The perception of control is pivotal in this calculus. When users feel they can manage their own information, their trust and willingness to engage with platforms are enhanced [28], a sentiment echoed by our participants' demands for granular data control.

A disconnect often exists between the privacy priorities of users and those of technical experts. Users typically focus on control over their personal data, a concern not always shared by expert evaluations that might prioritize technical vulnerabilities [14]. This underscores the need for user-centered privacy solutions. These concerns are not one-sided, as healthcare professionals also report significant anxieties about data security and the trustworthiness of telehealth technologies [37]. Such issues are intensified by the inherent power imbalances in the use of patient-generated health data [41]. Indeed, familiarity with a service and the involvement of a trusted care provider are key factors that motivate initial technology adoption [12].

The platforms themselves introduce further complications. Text-based interactions can be perceived as inefficient, prompting physicians to use conversational shortcuts that may compromise the quality of care [21]. While users pose a wide range of questions on these platforms [29], the underlying app ecosystem is often fraught with privacy risks. Many health apps lack transparent privacy policies [35], are plagued by insecure data handling that enables user profiling [11], and can cause unintentional harm [13]. Analyses of existing privacy policies confirm the urgent need for more robust data protection from both developers and policymakers [8].

In response, researchers have explored various system-level interventions. Some have studied how user interface design can inform users and address power asymmetries in online consultations [43], while others have focused on building tools like inter-doctor recommendation frameworks [23]. Although studies show conversational agents can be as effective as humans for preliminary information gathering [20, 22], they do not inherently resolve the privacy dilemma. Our work diverges significantly from these approaches. SafeShare directly confronts the privacy risks prevalent in the ecosystem [11, 35] by automating the strenuous "privacy labor" that we identified, thereby mitigating user burden and enhancing agency.

Other research has focused on privacy redaction. For instance, Zhang et al. [44, 46] and Albanese et al. [3] proposed frameworks for automatic text redaction and zero-shot sanitization respectively. However, our work is distinct in its focus on the unique medical context, where we explores how a localized LLM could achieve the balance between privacy and diagnostic utility.

3 Methodology

3.1 Study Setup

Participants and recruitment. This IRB-approved study recruited 12 participants (5 males, 7 females, age mean=22.2, SD=3.6) through distributing recruiting posters on online chat groups across a week. Participants were required to have at least one time of online consultation experience, however we did not require the total experience of participants as we aimed to recruit participants with diverse symptoms.

Interview process and analysis procedure. We conducted semi-structured interviews with participants to understand their information disclosure behavior on online medical consultation platforms, their cognition towards the information importance and the disclosure risk, their privacy protection behavior, user protection's (in-)effectiveness, the platform's behavior, their perception and the expectations. We conducted all interviews through online Tencent Meeting [1]. All interviews were recorded and transcribed after acquiring users' consent. We conducted thematic analysis on all

[1] https://meeting.tencent.com

interview results: two primary researchers separately coded 20% of the participants' interview scripts and formed the initial codebook. They then refined the codebook together and iteratively coded the rest of the interview scripts, as the interview was exploratory in nature. We also did not calculate the inter-rate reliability as a criteria because of the exploratory nature of the experiment and according to the previous guidance [9, 15, 31]. We reported the themes in the results section.

4 RQ1: Information Disclosure, Privacy Risks and Importance

4.1 Online Consultation Scenarios: How and Why

Our analysis analyzes how and why users engage with online medical consultations around two scenarios.

Low-stakes healthcare for triage and access. Users primarily leveraged online platform as a form of triage for low-acuity medical issues and to gain access to prescriptions. The consultations were often for minor ailments where a physical hospital visit was deemed unnecessary. These included common problems such as *"skin rashes, eye pain"* (P2) and *"colds, fevers, and coughs"* (P1), which were generally perceived as *"not very serious"* (P5). A second key use was prescription fulfillment, particularly for medications requiring a formal consultation step on e-commerce platforms before purchases, such as specific fever reducers or eye drops.

A trade-off calculus in platform selection. Users' decisions to use online or offline services were based on a deliberate weighing of convenience, accuracy and privacy. Online platforms were valued for their immediacy, with users noting, *"I can get a response very quickly"* (P12), a factor especially important for mitigating anxiety. Conversely, for conditions perceived as complex or serious, users defaulted to offline consultations. They values the diagnostic accuracy of physical examinations where a doctor *"can directly touch it and know what kind of nodule it is"* (P3), viewing remote assessments as potentially *"less accurate"* (P3). The anonymity of online platforms was also a significant affordance for sensitive health issues. As P6 noted, for conditions like *"HPV, HIV, it might be more acceptable to consult online"* (P6). However, this was balanced against concerns about data privacy. Some users were wary of platforms retaining their data or requiring them to photograph sensitive areas, preferring the ephemeral nature of an offline visit where *"each consultation is a one-time thing."* (P2) Finally, the economic model of a platform could also influence the choice. For instance, platforms that limit the number of follow-up questions per payment made offline visits more practical for the extended dialogue required for complex conditions.

4.2 Information Disclosure and the Privacy Calculus

We identified two primary findings, the scope of disclosure is highly dependent on the context of the medical interaction, and the privacy calculus that systematically distinguishes between medically necessary data and personally identifiable information.

4.2.1 The Scope of Disclosure is Context-Dependent. Participants did not have a static approach to sharing information but strategically adjusted the breadth and depth of what they disclosed based on the specific goal of consultation. We identified two distinct themes.

Formal disclosure for regulated transactions. When the purpose was to purchase a regulated item, such as a prescription drug, participants recognized the need for formal and verifiable disclosure. They consistently reported providing core identity details, including their name, age, and national ID number, along with their medical history. This was widely viewed not as an invasion of privacy but as a necessary procedural requirement. As one user reasoned, such verification is essential for platforms to prevent the misuse of controlled substances.

Holistic disclosure for diagnostic accuracy. For general diagnostic consultations, the scope of disclosure became significantly broader and more qualitative. To receive an accurate diagnosis, participants shared a holistic view of their condition and lifestyle. This included detailed accounts of *"recent conditions and habits"* (P1), visual evidence such as photos of *"small blisters on the finger"* (P11), and existing *"offline examination reports"* (P12). Participants even divulged seemingly tangential lifestyle details, such as a *"love for frozen or spicy food"* (P3) when consulting for a common cold, operating under the principle that more information would lead to a better diagnosis.

4.2.2 Privacy Calculus Distinguishes Medical From Personal Data. Participants performed a consistent mental trade-off to determine what was safe to share. This calculus was defined by two opposing but complementary considerations.

Willingness to share diagnostic information. The first privacy calculus principle is that information perceived as essential for diagnosis is shared willingly, even when acknowledged as private. Participants readily provided detailed symptom descriptions, medical histories, and revealing photographs because the utility of receiving an accurate diagnosis was deemed to outweigh the inherent privacy sensitivity of the data itself. This sentiment was perfectly captured by P5, who stated, *"[My condition] is also private information, [but] I am willing to provide this information to help my diagnosis."*

Resistance to sharing personal identifiers. The second and opposing principle is a strong resistance to disclosing PII, especially names and national ID numbers. This information was seen as the critical privacy boundary. The core fear was not the exposure of a medical condition in isolation, but the permanent, verifiable linking of that condition to their real-world identity. One participant powerfully illustrated this fear with the metaphor that connecting their name and ID to their health data *"would be like streaking"* (P11). While the necessity of providing an ID for regulated transactions was sometimes accepted as a form of *"social control"* (P7), the overwhelming consensus was that the ultimate privacy threat lies in the fusion of medical data with personal identifiers.

4.3 Perceived Privacy Risks

We identified three themes around the perceived risks associated with the conduct of both digital health platforms and individual medical practitioners.

Perceived risks of platform-level data exploitation. It captures participants' apprehension about how their information is collected, analyzed, and utilized by the platforms. A primary fear was the unauthorized commercial use of their sensitive health data. One participant articulated the risk of their data being used to *"recommend health supplements and drugs,"* which they believed *"could have an impact on users' health"* (P2). This anxiety was often substantiated by experiences of digital surveillance, with another user noting that after a single online consultation, *"every time I search for something, a pop-up window appears ... I feel like I've been recorded"* (P3). Beyond commercial exploitation, participants feared the creation of permanent and potentially damaging health profiles. For instance, a user expressed alarm that a one-time medication purchase for asthma-like symptoms had resulted in the platform permanently labeling them with an "asthma" diagnosis. They worried that *"this file will be shared ... and may affect their ability to purchase insurance in the future"* (P12). The basis for these fears was corroborated by an industry insider who confirmed that platforms systematically analyze user data, stating that *"chat records and communication records can all be heard ... we will use it to do some semantic analysis"* (P11).

Anxieties over unauthorized disclosure by individual practitioners. It relates to the conduct of doctors operating on the platforms. Distinct from the systemic risks posed by the platforms, users hold fear of personal data breaches stemming from individual actions. They were concerned about the potential for unprofessional behavior, as one user worried that a doctor might, *"our of morbid curiosity"*, take a screenshot of their confidential conversation and *"share it with others"* (P11). This highlights a specific vulnerability tied to the perceived integrity and professionalism of the individual doctors.

Distrust stemming from profit motives and public data display practices. It encapsulates a broader skepticism towards the healthcare platforms' business ethics. Participants expressed a general distrust of the commercial incentives driving these services, with one user bluntly stating that *"hospitals and platforms are profit-oriented,"* a concern they felt was especially pronounced with private healthcare entities (P9). This underlying profit motive was seen as a key driver for potential information misuse. Furthermore, this distrust was exacerbated by certain platform features, such as the practice of publishing anonymized medical records for public viewing as case studies. Even with the assistance of anonymization, the practice was a source of significant discomfort. As one participant explained, they would feel violated if their particularly *"outlandish"* case were to be shared publicly.

5 RQ2: Mitigation and Expectation

5.1 User-Initiated Mitigation Methods

Users described the landscape that, the burden of privacy protection falls largely on the user, with platform-provided measures being perceived as superficial and insufficient.

Proactive self-censorship and information control. With this most prevalent strategy, users employ tactics such as providing pseudonyms, offering an age range instead of a precise age, or strategically claiming the patient is a *"friend or relative"* to create psychological distance. It also includes the deliberate anonymization of visual data. When submitting images of symptoms, users are diligent about self-censorship, taking care to *"crop out the background"* or ensuring photos are taken against a plain backdrop. One participant stated they would *"definitely erase identifying information like my face or background details."* (P5) Similarly, when uploading existing medical reports, users would initially *"blur out personal information like their name or ID card number."*

Strategic platform selection based on perceived trustworthiness. Users gravitate towards services they perceive as reliable and secure. A clear preference was shown for the official applications of large, reputable public hospitals over third-party aggregator platforms. An industry-insider participant justified her exclusive use of a specific hospital's app by noting that on aggregator platforms, *"you have no way to guarantee the quality of the doctors,"* (P3) which she equated with higher privacy risks. This trust is also built on heuristic cues. One user favored a platform that originated from a *"professional medical forum,"* while actively avoiding others with *"chaotic interfaces and promotional ads."* (P9)

Perceived inadequacy of current platform measures. From the user's perspective, existing platform-level privacy features fail to build meaningful trust. Participants dismissed features like pop-up privacy agreements as superficial formalities that are seldom read or understood. While minor conveniences, such as one platform hiding the medicine name on the delivery package, were noted, they did little to address core data security concerns. The prevailing sentiment is one of profound skepticism. As one user articulated, *"They say they will protect our privacy, but we don't really know if they have."* (P11)

5.2 Challenges and Expectations for a Trustworthy System

The limitations of current measures give rise to significant challenges and a clear set of user expectations for reform. These are rooted in a foundational distrust of platform motives and a desire to reclaim control over their personal health data.

5.2.1 Core Challenges in Achieving Privacy. **Inherent limitations of user-driven mitigation.** While users believe their personal tactics are *"effective to a certain extent,"* they are acutely aware of the limitations. A fundamental challenge is the privacy-utility trade-off: effective diagnosis requires the disclosure of truthful and detailed medical information. As one user noted, *"if you don't say some things, they can't make a diagnosis"* (P6). This necessity often compels disclosure against their better judgment. This is compounded by a sense of learned helplessness, with some users operating under the assumption that their data has *"long been leaked"* through other channels, making extensive protection efforts feel futile.

Foundational distrust in the "black box" of platform operations. A significant barrier to trust is the opacity of platform data practices. For users, a platform's internal security mechanisms are a *"black box"* (P11), making it impossible to verify security claims. This distrust is structured around a tripartite model of trust: users generally trust doctors' professionalism (*"doctors have medical ethics"* (P10)), but this trust does not extend to the platforms, which are viewed as profit-driven entities. One user expressed this

dichotomy starkly: *"I trust their professional ability, but I default to assuming my information is being leaked [by the platform]"* (P12). This fear was validated by an insider who confirmed that platforms use consultation data to create a *"semantic database for analysis"* to improve their products, a practice described as *"dancing on a dangerous edge"* (P11).

5.2.2 User Expectations for System-Level Change. Fueled by these challenges, users articulated a clear vision for a trustworthy system, centered on demands for technical control and regulatory oversight.

Demand for granular user control and data ephemerality. This includes the ability to remain anonymous to the practitioner, with one participant wishing for a system where *"the doctor cannot see my personal data."* Users want the power to *"freely choose which parts of the medical record to display to the doctor."* A key component of control is the *"right to be forgotten."* Users desire the assurance that *"after the consultation ends, my personal information can be erased from their system"* (P2), viewing this as far more meaningful than a simple promise not to misuse data. The ideal, for some, is a system that severs the link to a permanent identity entirely, where *"I can just pay for the service without needing to log in or bind my phone number"* (P12).

Demand for enhanced transparency and accountability mechanisms A participant suggested the chat interface should feature a prominent *"claim stating that your chat history will only be used for a specific purpose and that screenshots or recordings are prohibited"* (P11). To enforce such policies, users expect platforms to implement *"very strict privacy training and security tests"* for doctors and to establish robust complaint channels and evaluation systems to hold both practitioners and platforms accountable for their conduct.

An imperative for external governance and regulation. Users expressed a profound lack of faith in corporate self-discipline and a corresponding demand for external oversight. This sentiment was captured unequivocally by one participant: *"I think the platform will never be self-disciplined ... only strong external constraints, like national policies, would make me feel at ease"* (P12)

6 SafeShare: A Data Protection Interactive Technique

To address the privacy tensions identified in our study, we designed SafeShare, an interactive technique that reframes data protection not as simple redaction, but as a process of contextual anonymization. This approach is designed around balancing the disclosure of medically relevant information with the robust protection of PII. SafeShare functions as an intelligent intermediary within the chat interface, leveraging a localized LLM to empower users with both the tools and the understanding to navigate the trade-off, thereby enhancing user agency and trust.

SafeShare comprises of a real-time anonymization and an in-context justification module. The real-time anonymization module automatically detects potential PII in both text and images according to pre-defined categories [47] elaborated in the prompts. Then the justification module, upon receiving the full set of identified entities, analyzes the user's query history to discern its specific diagnostic or information intent. Based on this intent, the module dynamically determines which identified entities are important for answering the query and which are not. This mechanism allows SafeShare to generate an optimal anonymization list for each specific context, describing which sensitive information to anonymize. It directly automates the strenuous and error-prone "privacy labor" and alleviating the cognitive load on the user.

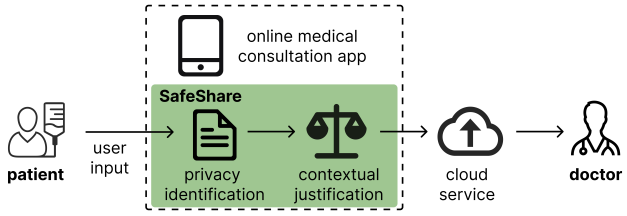

Figure 1: SafeShare acts as a bridge between users and cloud service, anonymizing medical private information.

After selectively anonymizing the inputs, SafeShare parses the output and replaces the users' original input locally, thereby protecting users' medical private information.

7 Evaluation of SafeShare

To validate the feasibility SafeShare, we conducted a quantitative evaluation, focusing on the performance of its core component: the real-time anonymization module. The objective was to assess the module's accuracy in identifying diverse types of PII from realistic medical text, which is critical for its ability to automate the "privacy labor" identified the interview.

7.1 Experimental Setup

Dataset. We utilized three different datasets related to online consultation. MedDG [26] is a large-scale, entity-centric Chinese medical dialogue dataset collected from Chinese online consultation platform *Doctor.Chunyu*[2], with 17,864 Chinese dialogues, 385,951 utterances. ReMeDi [40] contains 96,965 conversations and 843 types of diseases, including 1,557 conversations with fine-gained labels from *Doctor.Chunyu*[3]. IMCS21 [6] contains 4,116 samples, 10 diseases with 164,731 utterances from a Chinese online community, Muzhi[4].

Models and Metrics. We selected leading models with different sizes, brands and open or closed source status, including GPT-4o-mini from OpenAI, Deepseek-R1-7B from Deepseek, different model sizes of Qwen3 from Alibaba. We selected models with parameters smaller than 8B for evaluating on-device anonymization performance. The evaluation was conducted in a zero-shot setting, where each model was given a structured prompt defining the PII categories (e.g., NAME, ID) and instructed to extract all corresponding entities from the input text, balancing latency, cost and accuracy. The detailed anonymization prompts are shown in Appendix A.

7.2 Performance and Illustrative Case

Quantitative Results. The quantitative performance of the selected LLMs was evaluated on IMSC21, MedDG and ReMeDi. The

[2] https://www.chunyuyisheng.com
[3] https://www.chunyuyisheng.com
[4] http://muzhi.baidu.com

evaluation centered on both anonymization accuracy and appropriateness (see Appendix B for detailed definition and calculation of these metrics), which separately quantified the capability of correctly identifying sensitive information, and preserves the necessary clinical information for diagnosis. Table 1 and 2 showed the results. The findings indicated a clear performance trade-off among the models. *Qwen3-4b* model showed the highest anonymization accuracy across all three datasets, achieving scores of 89.64%, 84.86% and 82.41%, however exhibited lower scores in anonymization appropriateness.

Table 1: Anonymization accuracy of LLMs on different medical datasets.

Model \ Dataset	IMCS21	MedDG	ReMeDi
DeepSeek-R1-7B	77.96%	70.00%	75.52%
GPT-4o-mini	78.57%	72.77%	74.61%
Qwen3-1.7B	78.01%	73.22%	75.82%
Qwen3-4B	89.64%	84.86%	82.41%
Qwen3-8B	78.70%	87.04%	71.27%

Conversely, *Qwen3-1.7B* model showed the highest scores for anonymization appropriateness, with 92.40, 87.97 and 95.28 on IMCS21, MedDG, and Medical Dialogue datasets respectively out of 100. This suggests that while it was less precise in PII removal, it better preserves the diagnostic utility of the clinical text. Other models like *DeepSeek-R1-7B* and *GPT-4o-mini* achieved balanced performance, underscoring the potential of LLMs for balancing privacy and utility.

Table 2: Appropriateness of LLM's anonymization for diagnosing the symptom.

Model \ Dataset	IMCS21	MedDG	ReMeDi
DeepSeek-R1-7B	80.21	91.03	78.49
GPT-4o-mini	75.65	89.98	70.62
Qwen3-1.7B	92.40	87.97	95.28
Qwen3-4B	78.91	80.36	73.90
Qwen3-8B	70.04	76.03	68.44

Illustrative Case. To demonstrate the module's practical application, we presented a case of SafeShare, with inputs and output anonymization result, which contains multiple PII types[5]:

> *Original User Input:* "I am worried about the test results for my daughter Jane Doe from her appointment on May 20, 2025, with Dr. Smith at Peking University Hospital. We can be reached at 138-0000-0000 if needed."

SafeShare would present the following redacted version to the user for one-click approval before transmission:

> *SafeShare Anonymized Output:* "I am worried about the test results for my daughter [PATIENT] from her appointment on May [DATE], 2025, with [DOCTOR] at [HOSPITAL]. We can be reached at [PHONE] if needed."

This case illustrates the module's ability to correctly identify and appropriately anonymize PII entities within patients' description. SafeShare anonymizes sensitive information while retaining the meaning for the doctors to understand and diagnose.

[5]The original information is replaced for anonymization.

8 Discussions

Regulation compliance and anonymity. A core tension emerges between platform-enforced identification protocols and users' desire for anonymity. While some users understand the need for compliance, such as providing a real ID for purchasing controlled prescription drugs, as a necessary trade-off for societal safety, the general preference is for anonymous interaction [1]. This desire is particularly strong when dealing with stigmatized conditions where anonymity feels like a prerequisite for seeking care. However, users' attempts to maintain privacy through tactics like providing false information are often thwarted by system design. For instance, platforms can enforce compliance through mandatory real-name verification or by rejecting invalid ID numbers.

Feasibility of SafeShare. Our findings indicate a prevalent design failure in which communication platforms delegate the onus of privacy protection to users. This delegation compels individuals to engage in constant and effortful "privacy labor," such as manually redacting personal details from documents, cropping identifying features from images, and continuously assessing the necessity of information requests. Such practices impose a significant cognitive load on users, who may already be in a state of vulnerability. This model of user-managed privacy is not only burdensome but also inherently unreliable and unsustainable. We observed clear instances of "privacy fatigue" [7], where users ceased their redaction efforts due to the sheer effort involved.

To address these shortcomings, an effective approach would embed robust and usable privacy measures directly into the technique's design, thereby shifting the primary responsibility from the user to the platform. A technique like SafeShare could be integrated into commercial platforms that feature transparent communication channels [36]. Beyond merely detecting and anonymizing sensitive information, such a technique could also transparently communicate potential privacy risks to users [19, 45], therefore creating a supportive and secure environment.

Online, offline consultation and privacy risks. User privacy preferences and disclosure behaviors are highly contingent upon the specific medical condition [20]. Concerns range from the mandatory reporting of travel history for infectious diseases [24] and the fear of social stigma associated with conditions like HIV [4] and mental health issues [10], to a reluctance to share basic personal identifiers for common ailments [27]. This practice introduces a critical privacy challenge centered on data portability, as sensitive information is transferred from a trusted, regulated healthcare environment to a less secure digital one. This transition represents a significant point of vulnerability, underscoring the need for platform design that accounts for the broad, multi-modal health ecosystem.

Acknowledgments

This work was supported by the Natural Science Foundation of China under Grant No. 62472243 and 62132010.

References

[1] Na Young Ahn, Jun Eun Park, Dong Hoon Lee, and Paul C Hong. 2020. Balancing personal privacy and public safety during COVID-19: The case of South Korea. *Ieee Access* 8 (2020), 171325–171333.

[2] Ibrahim Al-Mahdi, Kathleen Gray, and Reeva Lederman. 2015. Online medical consultation: a review of literature and practice. In *Proceedings of the 8th Australasian workshop on health informatics and knowledge management*, Vol. 164. Australian Computer Society Sydney, 97–100.

[3] Federico Albanese, Daniel Ciolek, and Nicolas D'Ippolito. 2023. Text sanitization beyond specific domains: Zero-shot redaction & substitution with large language models. *arXiv preprint arXiv:2311.10785* (2023).

[4] Adrian Bussone, Simone Stumpf, and Jon Bird. 2016. Disclose-it-yourself: security and privacy for people living with HIV. In *CHI EA'16: Proceedings of the 2016 ACM annual conference on Human Factors in Computing Systems Extended Abstracts*. 1–4.

[5] Beizhesi Data Center. 2025. Overview of China's Online Doctor Consultation Industry: Market Size to Reach 57.59 Billion Yuan in 2024. https://caifuhao.eastmoney.com/news/20250529150848267627380.

[6] Wei Chen, Zhiwei Li, Hongyi Fang, Qianyuan Yao, Cheng Zhong, Jianye Hao, Qi Zhang, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. 2023. A benchmark for automatic medical consultation system: frameworks, tasks and datasets. *Bioinformatics* 39, 1 (2023), btac817.

[7] Hanbyul Choi, Jonghwa Park, and Yoonhyuk Jung. 2018. The role of privacy fatigue in online privacy behavior. *Computers in Human Behavior* 81 (2018), 42–51.

[8] Muhammad Hassan and Masooda Bashir. 2023. Unveiling privacy measures in mental health applications. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing*. 648–654.

[9] Stefan Albert Horstmann, Samuel Domiks, Marco Gutfleisch, Mindy Tran, Yasemin Acar, Veelasha Moonsamy, and Alena Naiakshina. 2024. "Those things are written by lawyers, and programmers are reading that." Mapping the Communication Gap Between Software Developers and Privacy Experts. *Proceedings on Privacy Enhancing Technologies* (2024).

[10] Yongquan Hu, Shuning Zhang, Ting Dang, Hong Jia, Flora D Salim, Wen Hu, and Aaron J Quigley. 2024. Exploring large-scale language models to evaluate eeg-based multimodal data for mental health. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 412–417.

[11] Leonardo Horn Iwaya, M Ali Babar, Awais Rashid, and Chamila Wijayarathna. 2023. On the privacy of mental health apps: An empirical investigation and its implications for app development. *Empirical Software Engineering* 28, 1 (2023), 2.

[12] Changmi Jung and Rema Padman. 2014. Virtualized healthcare delivery: understanding users and their usage patterns of online medical consultations. *International Journal of Medical Informatics* 83, 12 (2014), 901–914.

[13] Rachael M Kang and Tera L Reynolds. 2024. "This app said I had severe depression, and now I don't know what to do": the unintentional harms of mental health applications. In *Proceedings of the 2024 CHI conference on human factors in computing systems*. 1–17.

[14] Yi Xuan Khoo, Rachael M Kang, Tera L Reynolds, and Helena M Mentis. 2024. "That's Kind of Sus (picious)": The Comprehensiveness of Mental Health Application Users' Privacy and Security Concerns. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–16.

[15] Philip Klostermeyer, Sabrina Klivan, Sandra Höltervennhoff, Alexander Krause, Niklas Busch, and Sascha Fahl. 2024. Skipping the Security Side Quests: A Qualitative Study on Security Practices and Challenges in Game Development. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*. 2651–2665.

[16] Tarald O Kvålseth. 1989. Note on Cohen's kappa. *Psychological reports* 65, 1 (1989), 223–226.

[17] Robert S Laufer and Maxine Wolfe. 1977. Privacy as a concept and a social issue: A multidimensional developmental theory. *Journal of social Issues* 33, 3 (1977), 22–42.

[18] Law and Society Daily. 2023. Perfunctory Replies, Difficult Refunds, Prescriptions Before Consultation Leaking Privacy... Chaos in Online Health Consultation Platforms. *CCTV.com News* (7 June 2023). http://news.cctv.com/2023/06/07/ARTIA6sH2R05rYm8Q2Dk0193230607.shtml Original title: Chaos in Online Health Consultation Platforms.

[19] Hyunsoo Lee, Yugyeong Jung, Hei Yiu Law, Seolyeong Bae, and Uichin Lee. 2024. PriviAware: Exploring Data Visualization and Dynamic Privacy Control Support for Data Collection in Mobile Sensing Research. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–17.

[20] Brenna Li, Ofek Gross, Noah Crampton, Mamta Kapoor, Saba Tauseef, Mohit Jain, Khai N Truong, and Alex Mariakakis. 2024. Beyond the Waiting Room: Patient's Perspectives on the Conversational Nuances of Pre-Consultation Chatbots. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–24.

[21] Brenna Li, Tetyana Skoropad, Puneet Seth, Mohit Jain, Khai Truong, and Alex Mariakakis. 2023. Constraints and Workarounds to Support Clinical Consultations in Synchronous Text-based Platforms. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[22] Brenna Li, Saba Tauseef, Khai N Truong, and Alex Mariakakis. 2025. A Comparative Analysis of Information Gathering by Chatbots, Questionnaires, and Humans in Clinical Pre-Consultation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.

[23] Zhiying Li, Xiaonan Wu, Hongxun Jiang, and Xun Liang. 2025. Bridging Expertise: Doctor Recommendations for Cross-Disciplinary Collaborations in Online Medical Consultations. *Proceedings of the ACM on Human-Computer Interaction* 9, 2 (2025), 1–32.

[24] Jianqing Liu, Chi Zhang, Kaiping Xue, and Yuguang Fang. 2022. Privacy preservation in multi-cloud secure data fusion for infectious-disease analysis. *IEEE Transactions on Mobile Computing* 22, 7 (2022), 4212–4222.

[25] Shanshan Liu. 2021. Online Consultation: Privacy Leakage is Difficult to Prevent, Rights Protection Standards to be Determined. *People's Daily Online - Health and Life* (15 April 2021). http://health.people.com.cn/n1/2021/0415/c14739-32080340.html

[26] Wenge Liu, Jianheng Tang, Yi Cheng, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022. MedDG: an entity-centric medical consultation dataset for entity-aware medical dialogue generation. In *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 447–459.

[27] Zhihuang Liu, Ling Hu, Tongqing Zhou, Yonghao Tang, and Zhiping Cai. 2024. Prevalence Overshadows Concerns? Understanding Chinese Users' Privacy Awareness and Expectations Towards LLM-based Healthcare Consultation. In *2025 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, 92–92.

[28] Yumei Luo, Xiaoqi Li, and Qiongwei Ye. 2023. The impact of privacy calculus and trust on user information participation behavior in AI-based medical consultation-the moderating role of gender. *Journal of Electronic Commerce Research* 24, 1 (2023), 48–67.

[29] Xiaojuan Ma, Xinning Gui, Jiayue Fan, Mingqian Zhao, Yunan Chen, and Kai Zheng. 2018. Professional Medical Advice at your Fingertips: An empirical study of an online" Ask the Doctor" platform. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–22.

[30] Carlotta J Mayer, Julia Mahal, Daniela Geisel, Eva J Geiger, Elias Staatz, Maximilian Zappel, Seraina P Lerch, Johannes C Ehrenthal, Steffen Walter, and Beate Ditzen. 2024. User preferences and trust in hypothetical analog, digitalized and AI-based medical consultation scenarios: An online discrete choice survey. *Computers in Human Behavior* 161 (2024), 108419.

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

[32] Medical Arena. 2025. Cameras Installed in Gynecology, Doctor Dismissed After Removing Them. *Tencent News* (5 January 2025). http://news.qq.com/rain/a/20250105A0693X00 Published on the official account of Medical Arena in Shanghai..

[33] Tamir Mendel, Oded Nov, and Batia Wiesenfeld. 2024. Advice from a doctor or AI? Understanding willingness to disclose information through remote patient monitoring to receive health advice. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–34.

[34] New Hunan. 2023. Zhou Haimei's Medical Records Leaked, Two People Under Police Investigation, Their Identities Revealed! *New Hunan - The World* (14 December 2023). http://m.voc.com.cn/xhn/news/202312/17154238.html Hunan Daily New Media.

[35] Lisa Parker, Vanessa Halter, Tanya Karliychuk, and Quinn Grundy. 2019. How private is your mental health app data? An empirical study of mental health app privacy policies and practices. *International journal of law and psychiatry* 64 (2019), 198–204.

[36] Tian Shen, Yu Li, and Xi Chen. 2024. A Systematic Review of Online Medical Consultation Research. In *Healthcare*, Vol. 12. MDPI, 1687.

[37] Faiza Tazi, Josiah Dykstra, Prashanth Rajivan, and Sanchari Das. 2024. "We Have No Security Concerns": Understanding the Privacy-Security Nexus in Telehealth for Audiologists and Speech-Language Pathologists. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–20.

[38] The Beijing News. 2024. New Depth | Wang Chuqin's Privacy Exposed While Seeing a Doctor? What Responsibility Should Be Borne for Leaking a Patient's Medical Records? *Sina Finance* (20 December 2024). http://finance.sina.cn/2024-12-20/detail-ikywvvup7666270.d.html Official account of The Beijing News.

[39] Elisabeth Vodicka, Roanne Mejilla, Suzanne G Leveille, James D Ralston, Jonathan D Darer, Tom Delbanco, Jan Walker, Joann G Elmore, et al. 2013. Online access to doctors' notes: patient concerns about privacy. *Journal of medical*

Internet research 15, 9 (2013), e2670.
[40] Guojun Yan, Jiahuan Pei, Pengjie Ren, Zhaochun Ren, Xin Xin, Huasheng Liang, Maarten De Rijke, and Zhumin Chen. 2022. ReMeDi: Resources for multi-domain, multi-service, medical dialogues. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3013–3024.
[41] Dong Whi Yoo, Aditi Bhatnagar, Sindhu Kiranmai Ernala, Asra Ali, Michael L Birnbaum, Gregory D Abowd, and Munmun De Choudhury. 2023. Discussing social media during psychotherapy consultations: patient narratives and privacy implications. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–24.
[42] Ting Yu. 2020. Patient's Test Items Publicly "Announced" on Hospital's Big Screen? Hospital Claims It Was a System Malfunction. *Sichuan Online* (26 November 2020). http://sichuan.scol.com.cn/sczh/202011/57999884.html Source: West China City Daily. Editor: Liu Bo.
[43] Lanyun Zhang, Jiani Zhan, Verena Kwok Wai Wan, and Yanbin Wang. 2024. Designing and Evaluating Online Health Consultation Interfaces: A Perspective of Physician-Patient Power Asymmetry. *IEEE Access* (2024).
[44] Shuning Zhang, Lyumanshan Ye, Xin Yi, Jingyu Tang, Bo Shui, Haobin Xing, Pengfei Liu, and Hewu Li. 2024. " Ghost of the past": identifying and resolving privacy leakage from LLM's memory through proactive user interaction. *arXiv preprint arXiv:2410.14931* (2024).
[45] Shuning Zhang, Xin Yi, Shixuan Li, Haobin Xing, and Hewu Li. 2025. PrivCAPTCHA: Interactive CAPTCHA to Facilitate Effective Comprehension of APP Privacy Policy. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–20.
[46] Shuning Zhang, Xin Yi, Haobin Xing, Lyumanshan Ye, Yongquan Hu, and Hewu Li. 2024. Adanonymizer: Interactively Navigating and Balancing the Duality of Privacy and Output Performance in Human-LLM Interaction. *arXiv preprint arXiv:2410.15044* (2024).
[47] Jijie Zhou, Eryue Xu, Yaoyao Wu, and Tianshi Li. 2025. Rescriber: Smaller-LLM-Powered User-Led Data Minimization for LLM-Based Chatbots. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–28.

A Prompt Structure

We used LLM-powered anonymization because it could potentially understand the task, ensuring both privacy and utility. Besides, ensuring localized processing could minimize the data uploading, thereby protecting privacy and mitigating privacy risks. In the next subsections, we provided the design and implementation of the prompts, including the anonymization prompt and evaluation prompt.

A.1 Anonymization Prompt

The set of entity categories targeted for anonymization was established in accordance with prior privacy frameworks, encompassing: name, email, phone number, ID, online identity, geolocation, affiliation, demographic attributes, time, financial information, and educational records [47]. Different from prior work [47], to execute the NER task using a LLM, balancing diagnosis and anonymization, we engineered a structured prompt to precisely guide its behavior.

The prompt's architecture is multifaceted. First, it employs a role-playing instruction, assigning the model the persona of a *professional medical and privacy expert specializing in NER* to contextualize the task. Second, it defines the primary objective, which extends beyond simple entity extraction to enable a deliberate balance between privacy preservation and data utility. The model is tasked with identifying and extracting textual instances of sensitive entities whose sensitivity overweigh utility, from a given medical dialogue (input as *}dialogue_text{*) based on a predefined set of categories (input as *{entity_list_str}*). This identification is a prerequisite for subsequent anonymization, ensuring that non-sensitive, clinically relevant information remains intact, thereby preserving the utility of the data.

Third, it imposes a strict, machine-readable output schema that is critical for automated downstream processing. The prompt mandates a valid JSON output where keys correspond to the entity categories and values are lists of the exact textual excerpts. This structural requirement is reinforced with a one-shot example and includes instructions for handling null results (i.e., using an empty list or omitting the key) to ensure consistent and predictable model behavior.

A.2 Evaluation Prompt

To evaluate the dual objectives of effective PII removal and the preservation of clinical utility, we meticulously designed two prompt. The designs were necessary to operationalize the measurement of our two primary metrics, anonymization accuracy and anonymization appropriateness, while preventing task contamination and ensuring the validity of each measure. This approach allows for a focused and unbiased assessment of each objective independently.

The first prompt is designed exclusively to assess anonymization accuracy. To this end, we assign the LLM the persona of a privacy expert. The prompt provides the model with both the original, unaltered dialogue and the list of PII entities extracted by our system. The LLM is then tasked with a technical validation: to evaluate the correctness and completeness of the extracted entities against the source text one by one. We aggregated each assessment after using LLMs' evaluation.

The second prompt is engineered to evaluate anonymization appropriateness. Here, the LLM is assigned the different persona of a clinical physician. Critically, it is provided only with the fully anonymized version of the dialogue, with no access to the original text or the redacted PII. The model's task is to determine whether the remaining clinical information is sufficient to make a meaningful medical diagnosis, through providing a quantitative score.

The rationale for this two-prompt design is to create a controlled evaluation environment. By isolating the tasks, we ensure that the LLM's assessment of diagnostic utility (Prompt 2) is not biased by its knowledge of what specific PII was removed (Prompt 1). This separation is crucial for obtaining a reliable and objective measure of the delicate balance between privacy protection and data utility, providing a scalable and methodologically sound alternative to using human experts for each distinct evaluation task.

B Evaluation Metrics

To validate the effectiveness of the anonymization process, we employed two metrics: anonymization accuracy and anonymization appropriateness. Anonymization accuracy refers to the correctness of redacting private information, while anonymization appropriateness evaluates whether the anonymized answer retains sufficient information for an accurate patient diagnosis.

We estimated anonymization accuracy by using an advanced LLM (qwen-max) to judge the correctness of the entity recognition. To estimate anonymization appropriateness, the LLM was prompted to role-play as a physician and determine if it could accurately discern the patient's symptoms from the anonymized text.

Notably, to validate the accuracy metric, one experimenter manually coded a sample of 50 records. The inter-rater reliability between

the manual coding and the LLM's judgments, calculated using Cohen's Kappa [16], was 0.81, indicating substantial agreement. We decided to use an advanced LLM instead of human annotators because the dataset contains a wide variety of symptoms and disease categories, making it infeasibly complex to recruit physicians who are expert across all these different areas. We also acknowledge that future work could benefit from having human physicians perform the annotation task.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009