Intrusive by Default? Navigating Privacy by Design in LLM-Supported Care Documentation

Reinhard Kletter
TU Wien
Institute of Management Science
Theresianumgasse 27, 1040 Vienna
AT
www.tuwien.at
reinhard.kletter@tuwien.ac.at

Sabine Theresia Koeszegi
TU Wien
Institute of Management Science
Theresianumgasse 27, 1040 Vienna
AT
www.tuwien.at
sabine.koeszegi@tuwien.ac.at

In an LLM-based documentation support system for residential nursing homes, audio recordings of care worker–resident interactions are used to generate brief care reports. Drawing on ethnographic observations, we show that inconsistent privacy practices reflect a privacy calculus shaped by concerns about workplace surveillance. We examine the limits of operationalising Privacy by Design in a context where privacy concerns are intertwined with distrust toward both technology and the care organisation. As a technology probe to explore the technical potential for privacy protection, we developed an LLM-based privacy detection pipeline. While approximately three-quarters of pipeline-identified segments were confirmed as privacy-sensitive, unreliable speaker diarisation remains a major challenge. We argue that care workers should retain autonomy over their privacy preferences rather than relying on predefined technical safeguards. Our findings suggest that technical Privacy by Design measures alone are insufficient for establishing trustworthy AI documentation support systems in situated care work.

Privacy by Design, Care Documentation Support, Residential Nursing, LLMs, Value Sensitive Design

1. INTRODUCTION

Digital documentation is perceived as a burden by many care workers (Johnson et al. 2025). This challenge is particularly pronounced for second-language speakers, who often struggle with report writing, an issue further exacerbated by labor migration (Johannesen and Helleso 2019). Recent research has increasingly focused on automated support for doctor–patient conversations (Ng et al. 2025). However, these interactions are typically brief, structured and goal-oriented. In contrast, only a limited number of studies address automated support for nursing documentation (e.g., Hamdhana et al. 2024; Liu et al. 2025). Conversations in long-term care differ fundamentally from doctor–patient interactions. Nursing home residents often live in care facilities for extended periods, months or even years, resulting in sustained relationships with care workers that develop through repeated care encounters. Within these relationships, small talk plays a crucial role, often helping to ease the discomfort associated with care procedures (Macdonald 2016). Care-related communication frequently occurs alongside routine tasks rather than in dedicated conversational settings (Fleischer et al. 2009) and is embedded in everyday interactions rather than formal exchanges (Macdonald 2016). In long-term care, communication has a significant impact on residents' overall care experience (Shattell 2004). It is essential to deliver high-quality care (Sundler et al. 2020; Kwame and Petrucka 2021), to foster trust (Höglander et al. 2023) and ensure that residents feel heard and cared for (Macdonald 2016).

We deployed an LLM-based documentation support system (LDSS) in a residential nursing home. The system leverages automatic speech recognition and large language models (LLMs) to generate concise written reports from audio recordings of care interactions between care workers and residents and was developed following a value sensitive design (VSD) approach. In this study we draw on ethnographic accounts to capture the lived realities of privacy in care work. Based on these findings we propose a privacy detection pipeline

© Kletter et al.

as a technology probe, which integrates into the LDSS and identifies privacy-sensitive content within transcripts. In doing so, we address the limited body of work on the practical implementation of ethical approaches in technology development (Kumar et al. 2025). We apply the *Research through Design* approach and discuss our findings along the criteria for evaluating interaction design research. Based on these insights, we address the research questions: *"How is privacy in care documentation negotiated within care workers' situated work practices?"* (RQ1) and *"How can Privacy by Design be meaningfully integrated into LLM-based documentation support systems for care work"* (RQ2).

Our contributions are threefold. First, we critically analyse the lived privacy practices in the residential nursing context. Second, we advance methods for the automated detection of privacy-sensitive content in care worker–resident interaction transcripts. Third, we demonstrate that technical Privacy by Design (PbD) alone is insufficient to address persistent privacy concerns in real-world deployments.

2. RELATED WORK

The increasing adoption of LLMs in healthcare and long-term care underscores the importance of VSD and PbD as conceptual foundations. Related work in this field also highlights the dual role of LLMs in privacy research, both as a source of privacy risks and as a potential mechanism for mitigating them.

2.1. Value Sensitive Design and Privacy by Design in HCI

VSD conceptualises technology design as an explicitly value-oriented process, aiming to incorporate human values such as privacy, autonomy, and trust into sociotechnical systems (Friedman et al. 2013). Within human-computer interaction (HCI), VSD has become a foundational approach, prominently shaping discussions on ethics and design (Friedman et al. 2002). More recently, VSD has been extended to artificial intelligence (AI) contexts (Sadek et al. 2024; Umbrello and Van de Poel 2021).

PbD promotes a proactive integration of privacy into technological systems, emphasising privacy as a default rather than an afterthought (Cavoukian 2010). Central to PbD is the principle that privacy considerations should be embedded early in the design process and carried throughout the system lifecycle. In HCI, prior work has explored how to operationalise PbD in practice. For instance, Asgarinia (2023) proposes structuring privacy through hierarchies of values that can be translated into concrete design requirements. In the context of AI, Okon et al. (2024) highlights PbD's relevance for robust development, while Al Breiki and Mahmoud (2025) propose embedding PbD across the generative AI lifecycle, from data collection and storage to model design and deployment.

2.2. LLMs and Privacy

The literature on LLMs consistently emphasises their inherent privacy risks, particularly privacy leakage, and the potential extraction of sensitive training data (Chen et al. 2025). However, recent work also explores how LLMs themselves can be leveraged as tools for privacy and security. For instance, Grillaert et al. (2025) demonstrate how LLMs can support threat detection tasks in critical infrastructure contexts. PrivChatGPT, exemplifies this direction by integrating privacy-aware mechanisms into existing models to enhance data protection (Ullah et al. 2024). Similarly, the PrivacyAsst framework extends this idea to tool-augmented LLMs, focusing on safeguarding privacy in systems that interact with external tools and data sources (Zhang et al. 2024).

2.3. Automated Documentation in Healthcare and Long-Term Care

LLMs are increasingly used in healthcare, with applications including clinical text summarisation (Van Veen et al. 2024) and digital scribes that process clinician–patient conversations to generate clinical documentation (Van Buchem et al. 2021; Rotenstein et al. 2024; Ghatnekar et al. 2021; Ng et al. 2025). Commercial examples include Microsoft's DAX Copilot (Microsoft 2026), Ambience Healthcare (Ambience Healthcare 2026) and Abridge (Abridge AI 2026). Few approaches exist for nursing documentation support. These include extracting care-relevant information from telehealth interactions (Liu et al. 2025) and online consultations (Hamdhana et al. 2024), as well as organising nurses' dictations (Corbeil et al. 2025).

Long-term care presents distinct documentation requirements because care is organised around recurring routines defined in care plans (Keenan et al. 2008), while care worker–resident interactions also include relational and social conversation (Macdonald 2016). Existing tools to support in this context include voice-based nursing documentation (Voize 2026), LLM-supported structuring of residential care records (Vithanage et al. 2025) and mobile or wearable documentation support systems (Ehrler et al. 2021; Awan et al. 2018). Previous work combined automatic speech recognition and LLMs to extract care-relevant information from care worker–resident interactions (Kletter et al. 2026).

Although automated documentation may reduce note-writing time and documentation burden (Guo

et al. 2026), it introduces risks related to hallucinations, bias, and privacy (Kanaparthy et al. 2025; Ng et al. 2025; Bruff and Groves 2026). Proposed mitigations include locally deployed LLMs and synthetic data, although both have limitations (Jonnagaddala and Wong 2025; Tahera et al. 2026). Prior work proposed a framework based on contextual integrity theory (Nissenbaum 2004) to identify privacy-sensitive parts of care interactions (Kletter and Koeszegi 2026a). Beyond these technical and privacy risks, automated documentation systems may disrupt established care practices or create additional work for care workers (Wang et al. 2026; Seah et al. 2026). Their monitoring capabilities may also produce a sense of being surveilled (Petterson et al. 2024; Wang et al. 2026). Thus, preserving care workers' agency is essential when designing and introducing automated documentation technologies (Seah et al. 2026; Bruff and Groves 2026).

3. METHODOLOGY

While the broader longitudinal deployment examined the sociotechnical integration of the LDSS into care work practices and is reported elsewhere (Kletter and Koeszegi 2026b), the present paper focuses specifically on privacy aspects.

3.1. Research Approach

Building on ethnographic workplace studies in Computer-Supported Cooperative Work (CSCW) (Hughes et al. 1994; Jordan 1996; Randall and Rouncefield 2018) we observed care workers in their collaborative work with the LDSS during a 10-week longitudinal deployment in a residential nursing home in the Vienna metropolitan area, Austria. A 12-week preliminary study was conducted during which technical improvements were made to the system.

We draw on the notion of privacy as context-dependent, as articulated in contextual integrity theory (Nissenbaum 2004). However, as a culturally dependent concept (Bartoletti 2019), privacy has been conceptualised in various ways, ranging from multidimensional models (Burgoon 1982), to interaction regulation and boundary management (Altman 1975), to the simple right to be let alone (Brandeis and Warren 1890). Moreover, we distinguish between monitoring and surveillance, following the terminological divide in workplace observation research (Ball 2021). We use monitoring to denote the technical collection of work-related data for defined, legitimate purposes, consistent with its use in occupational psychology. In contrast, surveillance, as used in organisational sociology, emphasises asymmetries of power, control and discipline and refers to observation that is experienced, feared or intended as controlling and potentially sanctioning (Ball 2010; Ball and Margulis 2011).

Aiming to design the *right* thing, we use the *Research through Design* approach (Zimmerman et al. 2007) to evaluate our technical PbD implementation within the broader sociotechnical system in the discussion (Section 5).

3.2. Study Design

The LDSS was adapted for longitudinal deployment from our previous work (Kletter et al. 2026). It uses a local recording unit (a microphone with local storage) to capture audio from care worker–resident interactions during care activities. Care workers can upload these audio files via a separate upload unit to a secure LLM workstation. This separation of units is an intentional VSD decision aimed at preventing unintended uploads of audio files. The audio is automatically transcribed, and care-relevant information is extracted by an LLM. Following a VSD decision, all processing is performed with local LLMs to preserve privacy. Finally, care workers review and adapt the generated reports before entering them into their documentation software.

The privacy detection pipeline was introduced after completion of the longitudinal study as a technology probe (Hutchinson et al. 2003) exploring a potential technical implementation of PbD. The technical setup, including the used model, configurations, and prompting strategies, is detailed in the appendix (Section A). The probe was motivated by reported changes in residents' and care workers' behaviour when the LDSS recording unit was present, which were characterised by fewer personal exchanges (see Section 4.1). Because such informal interactions can provide valuable information and foster residents' comfort and trust (Macdonald 2016; Höglander et al. 2023), they should be preserved. To avoid including personal information in care reports, the pipeline identifies transcript segments that fall within predefined categories of privacy-sensitive information (see Section A.2) and filters them according to a set of rules before the transcripts are used for report generation. The pipeline uses only local LLMs on limited computing resources. Figure 1 illustrates how it integrates conceptually into the LDSS.

To evaluate the pipeline, a ground truth for all LLM-selected segments was established through human annotation. One researcher annotated each conversation segment to determine whether it contained privacy-sensitive information. For each candidate, the annotator indicated agreement or disagreement with the LLM's rationale for classifying

it as privacy-sensitive. Where necessary, the annotator also corrected the identified disclosing speaker, the affected individual and the assigned privacy category.

3.3. Recruitment and Participation

Participation in the longitudinal study was voluntary and counted as work time for the care workers. Throughout this paper, *care worker* refers to staff providing direct resident care, including different professional roles. Care workers were recommended by the care organisation and approached individually by the researchers, who explained the study details and obtained informed consent for participation. Participating residents were similarly recommended by the care organisation and care workers and approached individually by the researchers. Residents were only recruited if they were able to fully understand the study and provide informed consent. All care interaction recordings were made by the participating care workers only with residents who had given their consent and had been informed about the recording. Four care workers expressed interest in participating in the longitudinal study. Of these, Barbara, Natalia and Ana were interviewed. Natalia used the recording unit but did not submit any recordings for further processing, whereas Barbara and Ana used the full system. The fourth interested care worker, Monika, did not sign the consent forms required for study participation. In addition, two nursing home managers (Sandra and Julia) participated in an informal stakeholder conversation.

3.4. Data

In the longitudinal study, we conducted seven semi-structured interviews with three care workers. Barbara and Ana were interviewed approximately 2, 5/6, and 10 weeks into the deployment, whereas Natalia participated in only one interview. All interviews were audio-recorded, transcribed and analysed in German; quotations were translated into English for publication. The interviews addressed documentation practices, the integration of the system into existing workflows, its perceived utility, and perceptions of documentation quality (see Kletter and Koeszegi 2026b). In the final interview round, we focused particularly on barriers to adoption, reasons for non-use, and participants' reflections on both the system and the study process. The informal stakeholder conversation with the two nursing home managers was also recorded. Field notes were taken during each site visit and subsequently expanded to capture observations and initial analytic reflections. These qualitative materials were considered alongside digital traces of the system use. Across the preliminary and main studies, the two care workers submitted 27 recordings involving six unique residents for processing and evaluated 18 system-generated reports. Details on usage patterns, including the reasons for non-use and symbolic use, are reported elsewhere (see Kletter and Koeszegi 2026b).

A separate dataset of 60 transcripts of care worker–resident interactions was used to evaluate the privacy detection pipeline. Each transcript contained approximately 200 lines of dialogue. Collectively, the transcripts involved eight care workers and ten residents, all of whom had consented to participate and were aware that the interactions were being recorded. Of the 60 transcripts, 37 originated from the testing and evaluation phases reported in our earlier study (Kletter et al. 2026). The remaining 23 were collected during the longitudinal deployment (see Kletter and Koeszegi 2026b) and contained automatically generated speaker labels (e.g. *SPEAKER_00*) at the beginning of each line, whereas the other transcripts did not identify individual speakers. However, automated speaker diarisation proved unreliable in this setting: care workers typically dominated the conversation and were recorded more clearly because they wore the microphone, while residents often spoke softly or hesitantly, making their contributions difficult to understand even for human listeners.

3.5. Data Analysis

While the broader ethnographic study examined the integration of the LDSS into everyday work practices, this work focuses specifically on *privacy-related tensions*. We therefore conducted a directed qualitative content analysis (Hsieh and Shannon 2005) of the interviews from the longitudinal study, building on the coding framework developed in the broader study (Kletter and Koeszegi 2026b). Two researchers with backgrounds in the technical and social sciences independently coded the material using this coding framework. The coding scheme was iteratively refined to incorporate inductively emerging codes. The final coding scheme comprised four overarching categories (example codes in parentheses): (1) Privacy and data governance (care worker and resident privacy, surveillance, data access, trust, audio recording concerns); (2) Autonomy and agency (responsibility, human oversight, accountability); (3) Power relations (study participation, decision-making authority, management attitudes); and (4) Technology attitudes (positive attitudes, negative attitudes, AI attitudes, job security). Field notes were used to contextualise interview accounts and identify recurring tensions and contradictions between reported experiences and observed work practices. Throughout the analysis, the researchers held reflexive meetings to discuss coding decisions,

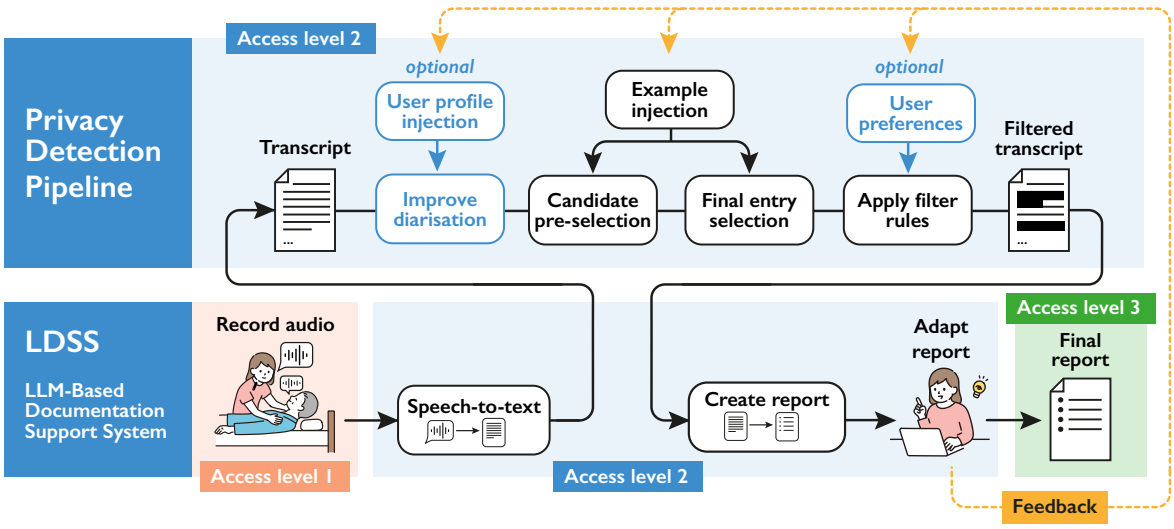

Figure 1: *Integration of the privacy detection pipeline within the LDSS. Different access levels are indicated and a feedback loop is suggested to improve detection quality. Vector icons used are by Soco St in CC Attribution License via SVG Repo.*

validate interpretations, and resolve disagreements through consensus.

3.6. Use of AI Tools

AI-based tools were used to support parts of this work. ChatGPT was used to assist in writing code for both the system and the data analysis, including generating system and user prompt suggestions, and to support refinement of the paper's thematic focus. For language editing, DeepL and ChatGPT were used to improve readability and style, and to correct grammar, spelling, punctuation, and tone in text written by the authors. Perplexity was used in addition to conventional academic databases to support literature discovery. All outputs generated by these tools were reviewed, validated, and, where necessary, revised by the authors, who take full responsibility for the content of this publication.

4. RESULTS

4.1. Privacy in Care Documentation Practice

Although the LDSS was developed and tested with care worker involvement and incorporated several privacy-related VSD decisions, resistance to participating in the longitudinal study nevertheless emerged, primarily because of privacy concerns about the potential use of audio recordings for workplace surveillance by the care organisation.

> *"Because back then, everyone refused [to participate in the study]. Yes, exactly those concerns, that someone will be checking up on me."* (Barbara)

Participants explained that the stressful work environment can lead to mistakes or inaccurate statements during care interactions that might later be used against them.

> *"There are days when some residents are so restless. And what if I ever scream without realising it? Will he [the system] record it and call the police, or what?"* (Ana)

Recording one's own voice was perceived as something intimate and permanent that cannot be changed. Shame stemming from a lack of language skills while being recorded also played a role. The language barrier also contributed to Monika's uncertainty about participating in the study, which could not be resolved until the very end despite three informational meetings. Although the participant information was deliberately kept simple, it still appeared to be overwhelming.

> *"I think it's difficult, partly because of the language. They [other care workers] remain in this position, where they're safe, and anything else is a 'no, thanks'."* (Ana)

Fears of workplace surveillance are an issue that the nursing home management is well aware of, since it affects other areas as well.

> *"But you often see that in surveys, too. Our employees are often afraid because they think, 'Okay, who's going to see this now?'"* (Sandra, Manager)

While the management clearly opposes such surveillance and acknowledges that employees' language barriers play a role, they are puzzled as to why employees have such strong reservations.

> *"Or what would be so bad if I were to listen to it? Which I would never do, but what if I were to listen to it? What would I hear?"* (Julia, Manager)

At the same time, participating care workers reported that some colleagues had recently started using generative AI such as ChatGPT to assist their written reports. They reportedly either dictate or write

contents in their native language and let the LLM translate and structure it for their report. This is in addition to other third-party tools already in use, such as translators and grammar checkers.

> *"I didn't know this before, but I thought, 'My God, she barely speaks German.' Yes, now she writes reports that you can not believe, yes."* (Barbara)

This behaviour may appear paradoxical. Although the LDSS is tailored to their use case and incorporates multiple privacy-preserving measures, care workers resist it because it relies on audio recordings. At the same time, care workers willingly share sensitive data, including their voices, with third-party applications, even though data processing and use remain opaque. This practice also raises questions about human autonomy and accountability for the reports. Because the tools are not designed for this purpose, they may produce reports that are difficult for both their authors and other care workers to understand.

> *"ChatGPT and Google don't always write things the way we're supposed to. They might phrase things differently, and sometimes it just doesn't fit at all. I sometimes read reports where I have to wonder, 'What was she trying to say there?' And you can tell those aren't her words, because you know that person."* (Natalia)

However, care workers' choice of technological support suggests that they prioritised avoiding workplace surveillance over the privacy risks posed by third-party tools or, alternatively, that they perceived themselves as sufficiently anonymous when using such services. The employer's capacity to sanction them appeared more immediate and consequential than the comparatively abstract risks associated with third-party data practices. In interviews, care workers reported that there was similar initial resistance toward video-based fall prevention cameras as toward the LDSS. These cameras were introduced nonetheless, and the concerns have reportedly since faded into obscurity.

Even those who used the LDSS during the longitudinal deployment exhibited privacy-enhancing behaviours. Participants reported bringing the recording device only to residents who had consented to study participation, thereby avoiding unintended recordings. However, the recording unit allowed recordings to be deleted, and unintended uploads were impossible because the device operated entirely offline. Also, when participating residents opted out of being recorded, care workers removed the recording unit from the room rather than simply refraining from activating it.

> *"A resident [...] wanted to have a private conversation with me. [...] I took the device outside because I wasn't sure if I'd turned it off properly. That makes me feel safer."* (Natalia)

This tendency to err on the side of caution, despite repeated assurances that unintended uploads were technically impossible, suggests a deep-seated distrust of the technology. At the same time, it reflects a strong commitment to protecting residents' privacy, a concern also expressed on other occasions.

> *"But if everything that's recorded goes straight into our documentation software, maybe that brings up the fear: 'Well, what if I say something wrong today?' Or what if a resident says something she doesn't want [to be recorded]?"* (Ana)

In contrast to these privacy-aware practices, we also observed that computers used for documentation were frequently left unlocked, thereby exposing residents' data. Moreover, although a dictation feature was implemented as an alternative to recording during care activities, allowing care workers to dictate only relevant, non-privacy-invasive information, it was not adopted by the participating care workers. Generally comfortable with being recorded, they preferred automatically generated reports from natural care interactions over the additional effort required for dictation. It was also noted that dictation would require a quiet room and that written notes beforehand would be helpful.

> *"I need a script. For me, you know, I can't speak that freely. So I prefer it this way."* (Barbara)

Participating care workers expressed a preference for merging the LDSS recording and upload units to reduce additional steps in system use, thus actively asking to revise this deliberately privacy-preserving design decision in favor of usability. Moreover, they reported that both they and the residents adapted their communication and behaviour when using the audio recording device: conversations became more formal, and less private information was shared, thereby compromising the integrity of the interaction.

> *"Well, like I said, it's a little weird for the residents too, yeah. Because it's a bit of an intrusion on their privacy. And they just talk differently then."* (Ana)

However, Ana also offered to show residents the edited LLM-generated reports and assured them that private conversations would not be included.

4.2. Technology Attitudes

Beyond the privacy concerns, the interviews also revealed broader, generally critical attitudes toward new technologies, especially AI.

> *"This whole thing really scares me. I mean, artificial intelligence in general."* (Barbara)

A distinct concern, analytically separate from privacy, was the loss of decisional autonomy to future technological tools: at stake is not who may access one's data, but who ends up making one's decisions.

> *"Somehow, everything has gotten too fast. Slowly, it's getting out of our hands. Eventually, someone else will decide, or a device will decide, what I want or don't want."* (Ana)

Moreover, surveillance is perceived as ubiquitous and unavoidable, indicating a sense of resignation.

> *"Yeah, that's already the case in this world anyway. We're under surveillance everywhere, in one way or another."* (Barbara)

However, attitudes were mixed. Although participating care workers expressed fears about emerging technologies, they were also proud to contribute to the development of the LDSS, which uses AI.

> *"Like I said, I've always been afraid of things like that, and computers in general are like black magic to me. But even so, I'm taking part in a project like this, which actually makes me proud."* (Barbara)

Participants emphasised the LDSS's potential benefit for future care workers, while noting it would take time to become accustomed to it.

4.3. Distinguishing Private Information

The lack of reliable speaker diarisation frequently led to the inclusion of private care worker content in LDSS-generated reports, raising several challenges. First, care workers' privacy may be compromised when personal information is inadvertently incorporated into care documentation. Second, the integrity of resident records is undermined if such content is not reliably removed. Third, care workers must invest additional effort to distinguish irrelevant personal content from care-relevant information, a process they described as both tedious and detrimental to the overall utility of the LDSS. In addition to preserving the integrity of care interactions, these findings motivated the technology probe of the privacy detection pipeline, which aims to filter privacy-sensitive content at the transcript stage, before reports are generated, according to the rules outlined in Table 1.

The rationale for the rules is informed by the accounts of care workers in this and prior studies (Kletter et al. 2026; Kletter and Koeszegi 2026a). Information concerning the resident's physical and psychological health is relevant for care documentation and should therefore be retained (R1). Information about acquaintances, appointments, family, the resident's thoughts, and other personal aspects of the resident's life may be relevant for biographical work and should likewise be retained (R2). As care workers review and adapt the generated reports, they can manually remove information deemed irrelevant. In contrast, deeper personal values and beliefs of residents should not be included in formal documentation (R3). Privacy-sensitive information concerning care workers or third parties (others) must never be included in the documentation (R4 and R5). For the operationalisation the pipeline must thus be able to identify the privacy category of the shared information and the affected individual (subject).

4.4. Automated Privacy Detection

4.4.1. Privacy-Sensitivity and Classification

The privacy detection pipeline classified 394 of 397 pre-selected candidates as privacy-sensitive (mean confidence: 73.5%; see Sections A.2 and A.3 for details of the technical setup). Human annotation confirmed 287 of these as privacy-sensitive, corresponding to a false-positive rate of 27.2%. All three rejected candidates were confirmed as non-sensitive, resulting in no false negatives within the candidate set. The pipeline's reasoning for classifying content as privacy-sensitive was judged acceptable in 87.1% of confirmed cases; errors most commonly involved misidentifying the affected individual.

Figure 2 compares the AI-predicted and human-annotated privacy categories for the 287 privacy-sensitive entries. The diagonal cells indicate correct classifications, while off-diagonal cells show misclassifications. Each row-normalised cell shows its proportion and count, with darker shading indicating higher proportions. Category classification was strongest for *health* (96%), *family* (85%), and *values* (83%), although the latter occurred only six times. Most *acquaintance* entries were classified as *family*, and *appointments* were frequently classified as *health*, reflecting ambiguities that are difficult to resolve from isolated transcript excerpts.

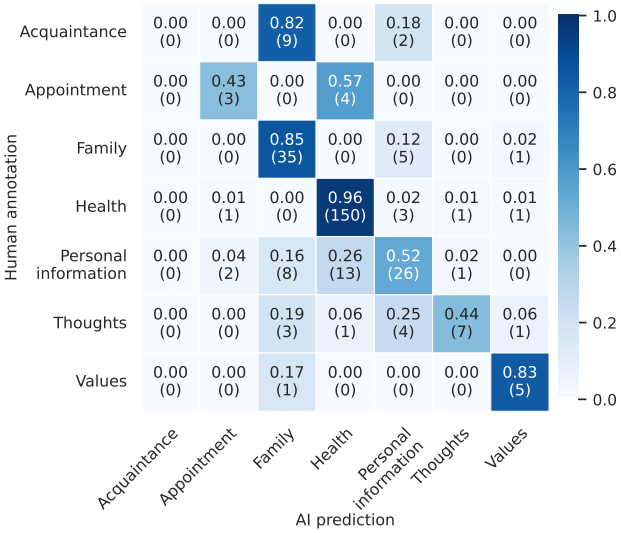

Figure 2: *Confusion matrix of AI-predicted versus human-annotated privacy categories (privacy-sensitive entries only).*

The pipeline reliably identified residents' disclosures of privacy-sensitive information (95% accuracy) but misclassified 60% of care worker disclosures, despite care workers disclosing sensitive information more frequently than residents (187 vs. 98 instances). The pipeline showed a strong bias towards identifying residents as the affected individual, resulting in 76 misclassifications and 211 correct

Table 1: *Rules for operationalising privacy in nursing documentation. R1–R2: non-filtering rules; R3–R5: filtering rules.*

Rule	Condition	System Decision
R1	The subject is the **care recipient** and the information concerns *health*.	Care-relevant
R2	The subject is the **care recipient** and the information concerns *acquaintances*, *appointments*, *family*, *personal information*, or *thoughts*.	Maybe care-relevant
R3	The subject is the **care recipient** and the information concerns their *values*.	Not care-relevant
R4	The subject is the **care worker** or *both*.	Not care-relevant
R5	The subject is a *third party* (*other*).	Not care-relevant

classifications, of which 210 involved residents and only one involved a care worker.

4.4.2. Rule-Based Filtering

We applied the filtering rules (see Table 1) to the 394 AI-classified privacy-sensitive entries and compared the results with the human-corrected ground truth. R1 was applied 226 times, with 17 errors (7.5%); R2 152 times, with 58 errors (38.2%); R3 11 times, with five errors (45.5%); R4 three times, with two errors (66.7%); and R5 twice, both incorrectly (100%). Overall, 84 entries were handled incorrectly (21.3%). The filtering rules R3–R5 incorrectly removed non-sensitive content in 9 of 16 cases (error rate: 56.3%). In contrast, the non-filtering rules R1–R2 incorrectly retained sensitive content in 75 of 378 cases (error rate: 19.8%). However, false-positive AI classifications had a desirable effect in R1–R2 because the retained entries were not actually privacy-sensitive. Most critically, owing to unreliable speaker diarisation, only three of the 58 privacy-sensitive instances affecting care workers were automatically filtered, yielding an error rate of 94.8%.

5. DISCUSSION

We analyse our contribution through the four *Research through Design* criteria defined by Zimmerman et al. (2007): process, invention, relevance, and extensibility (Sections 5.1 to 5.4).

5.1. Process

5.1.1. Challenging Researchers' VSD Decisions

Explanations that recording and uploading had been deliberately separated to prevent unintended uploads were largely disregarded. Rather than relying on this technical safeguard and communicating it to residents, care workers removed the recording unit from the room, indicating considerable distrust of the LDSS technology. This aligns with their broader critical attitudes toward technology. At the same time, removing the device allowed care workers to visibly acknowledge and respect residents' privacy preferences. The device's visibility and physical presence thus made otherwise invisible documentation practices tangible, even though care-relevant information from interactions may later be documented regardless of whether the device is present. Despite its privacy-preserving intent, the separation of the recording and upload units contributed to infrequent use. Care workers preferred a combined unit requiring fewer interaction steps. This suggests that privacy-preserving measures should not disrupt established work practices. By contrast, other VSD-informed decisions that introduced no additional friction, such as local data processing, were perceived neither positively nor negatively by care workers.

5.1.2. Intrusive vs. Trustworthy Technology

The study highlighted that privacy risks associated with the LDSS were not judged objectively, but rather relative to the existing status quo (Adjerid et al. 2018). Care workers focused on whether it could introduce new privacy risks by enabling workplace surveillance, revealing a conceptual mismatch between management and care workers: while management viewed the system's monitoring capabilities as an incidental technical by-product that would not be used, care workers interpreted the same capabilities as creating the possibility for surveillance in the form of asymmetric observation that could ultimately be used against them (Ball 2010; Ball and Margulis 2011). Addressing this tension and fostering trust in the technology require more than technical privacy protections: the system's capabilities and limitations must be clearly communicated, and credible governance measures must be established.

5.1.3. Embedding Privacy

Privacy risks arise not from technical systems alone but from their entanglement with situated work practices, such as inadvertently exposing resident data. Accordingly, privacy safeguards should align with everyday documentation workflows to avoid encouraging workarounds that undermine their intended purpose. For example, automatic logout was introduced because computers used for documentation were frequently left unlocked. However, when care workers interrupt documentation to attend to a resident, logging back into the virtual desktop environment is time-consuming. To avoid repeated logins, they may intentionally leave computers unlocked and attempt to return before the session expires.

5.2. Invention

The privacy detection pipeline presents a novel approach to embedding privacy protection within an LDSS. Approximately three-quarters of pipeline-identified segments were correctly classified as privacy-sensitive, and agreement on the *health* category was particularly high. However, the lack of reliable speaker diarisation frequently led to the misattribution of statements to residents and the incorrect assumption that the residents' privacy was affected, thereby introducing substantial errors when automatically applying filtering rules. From a technical perspective, the probe therefore highlighted the need for improved speaker diarisation. However, we acknowledge privacy is a complex social issue that cannot be fully resolved through technology-driven PbD approaches alone (Danezis et al. 2015; Gürses 2014), but instead requires accompanying organisational processes (Balboni and Macenaite 2013). Our findings reinforce this perspective and further challenge AI-solutionist assumptions that most problems can be solved technologically (Reyes-Cruz et al. 2025). However, as others have argued, minor privacy violations can be mitigated through technical solutions (Danezis et al. 2015), making the pipeline an important additional technical safeguard.

5.2.1. Access Controls

To mitigate surveillance concerns, we propose three access levels with clearly communicated system boundaries (see Figure 1), in line with related work (Asgarinia 2023). Raw recordings should remain local and accessible only to the respective care worker (*access level 1*), even if the recording and upload units are combined. Uploaded audio and the resulting unfiltered and filtered transcripts should be accessible only to the care worker and researchers or, in practice, the service provider during processing (*access level 2*); audio files should be deleted by default once processing is complete. The care organisation should receive only the final report reviewed and adapted by the care worker (*access level 3*).

5.2.2. Autonomy and Feedback Mechanisms

Long-term care facilities' institutional goals, including documentation requirements, can conflict with residents' desire for privacy (Tufford et al. 2018) and automated documentation of care worker–resident dialogues may intensify this tension. Because the boundary between care-relevant and other information is often blurred, care workers may manually exclude care-relevant, but particularly sensitive details, from formal records (Kletter et al. 2026). Documentation should therefore remain only partially automated. Care workers' autonomy should be preserved by enabling them to review and adapt reports and to delete audio, transcripts and associated user-profile data (Seah et al. 2026; Bruff and Groves 2026). Moreover, inspired by the "ethical knob" proposed for self-driving cars (Contissa et al. 2017), we propose a configurable "privacy knob" for the privacy detection pipeline. Care workers could adjust the LLM-calculated confidence threshold at which information is classified as privacy-sensitive or disable filtering rules, balancing the risks of disclosing private information and omitting relevant content from LLM-generated reports. Transparent settings should be complemented by continuous feedback mechanisms that allow care workers to rate or flag entries and identify privacy-sensitive or missing information. This feedback could be traced back to the corresponding transcript segments to generate additional training examples, refine user profiles and improve subsequent filtering (see Figure 1). Building on LLM personalisation (Zhang et al. 2025), these user profiles could also support an alternative approach to speaker identification based on individual user characteristics.

5.3. Relevance

5.3.1. Privacy Paradox vs. Calculus

In practice, privacy-infringing products may even achieve greater success (Danezis et al. 2015), raising questions about the practical relevance of implementing PbD in the LDSS context. Despite our system's grounding in VSD principles, care workers who resisted its use instead turned to tools such as ChatGPT, dictating, and thus sharing their voices, a particularly intimate form of personal data, with third-party services. At first glance, this behaviour exemplifies the privacy paradox: the discrepancy between individuals' stated privacy concerns and their actual disclosure practices (Kokolakis 2017). However, read through the lens of a privacy calculus (Dinev and Hart 2006), this behaviour appears less paradoxical than selective: the employer constitutes a concrete, proximate audience with sanctioning power, whereas the data practices of third-party providers remain an abstract risk. This is consistent with research showing that concerns about identifiable, socially or hierarchically close audiences tend to outweigh institutional privacy concerns about how technology providers process one's data (Raynes-Goldie 2010). Moreover, this highlights that the perceived usefulness strongly shapes the acceptance of privacy trade-offs (Distler et al. 2020).

5.3.2. Operationalisation of PbD

Given the abundance of AI-based documentation support systems (see Section 2.3), operationalisable PbD approaches are increasingly important. Unlike imposed video-based fall-prevention systems, to which care workers eventually became accustomed, technologies should deliver their benefits without compromising privacy preferences, consistent with

the positive-sum principle of PbD (Cavoukian 2010). Building trust and addressing privacy concerns are essential to technology acceptance (Dhagarra et al. 2020), whereas value-insensitive designs may provoke resistance and worsen attitudes toward technology. Indeed, even participating care workers expressed technological fatalism resembling what prior work describes as privacy cynicism or digital resignation (Hoffmann et al. 2016; Draper and Turow 2019). The technology probe showed that, although LLMs introduce inherent privacy risks, they may also enable a positive-sum outcome. By filtering sensitive information, the pipeline could protect residents' and care workers' privacy, maintain the integrity of interactions, reduce post-editing effort, and preserve the integrity of residents' data records.

5.4. Extensibility

Assurances of research independence and technical VSD-decisions did not alleviate privacy concerns. Designers and data holders should instead address such concerns proactively and make them central to communication (van Rest et al. 2012). Workshops involving a critical public of diverse stakeholders can promote adoption, demonstrate system utility, invite critique, refine countermeasures, and build trust with care workers (Ayalon and Toch 2021; Boaz et al. 2018; Kaisler and Missbach 2020). Given concerns about workplace surveillance, these workshops must account for power imbalances in participant selection and enable care workers to express their preferences openly (Wacnik et al. 2025). Although derived from residential long-term care, our findings suggest that privacy in other research settings must likewise be negotiated through social practices, such as participatory activities and trust-building, organisational policies and technical measures.

5.5. Limitations and Future Work

Methodological and ethical constraints limited the collection of real-world care interactions. Few care workers joined the longitudinal study, and consent requirements excluded residents unable to understand its nature and implications, an important limitation given the prevalence of dementia in residential care. The resulting dataset therefore comprises few transcripts from a small participant pool. Moreover, because privacy is culturally situated, its classification is shaped by the backgrounds of annotators, care workers, and residents. As the pipeline evaluation relied on a single human annotator, the ground truth does not capture inter-annotator variation. Thus the findings should not be broadly generalised.

Future Work. Given the limited evidence that operationalising PbD can alleviate deep-seated concerns about intrusive technologies, future work should examine how its principles can be implemented and communicated through organisational, social and technical measures. The LLM's strong bias towards identifying residents as the affected individual should be investigated. Personalising the LLM to identify speakers through user profiles, rather than relying on automated speaker diarisation, is a promising research direction. However, leveraging more privacy-sensitive information internally to enable privacy-aware outcomes must be accompanied by appropriate technical and organisational safeguards. Future evaluations should incorporate user feedback mechanisms and manual transcript reviews to identify missed privacy-sensitive content. They should also assess whether filtering removes context needed for report generation, whether generated reports impact care workers' reflections on residents' conditions, and how partly automated documentation affects residents' privacy.

6. CONCLUSION

Our ethnographic observations reveal a discrepancy between expressed and enacted privacy practices. Care workers adopted third-party tools while declining to participate in a study using a VSD-informed prototype because of concerns about workplace surveillance in future deployments. This behaviour reflects a privacy calculus rather than a privacy paradox. To investigate the operationalisation of PbD, we implemented a privacy detection pipeline as a technology probe. Using a locally deployed LLM that operates on limited computing resources, the pipeline detects and classifies privacy-sensitive information in care worker–resident transcripts. Approximately three-quarters of pipeline-identified segments were confirmed as privacy-sensitive by human annotation. However, unreliable speaker identification frequently led to misclassification of the affected individual. Feedback mechanisms and speaker identification through personalisation could improve performance. Our findings show that technical PbD measures alone cannot alleviate deeply rooted privacy concerns. Trust in LDSSs also requires clear communication, appropriate governance, care worker autonomy, and active stakeholder involvement.

ACKNOWLEDGMENTS

This work was funded by a FWF #ConnectingMinds grant to the project Caring Robots // Robotic Care (CM 100-N). The research proposal was peer reviewed by the ethics committee at TU Wien, under case number 035_12052023_TUWREC. We would like to thank all the participants in this study and Alexander Witzler for his assistance with the interviews and data coding.

REFERENCES

Abridge AI (2026). Abridge AI. Available from https://www.abridge.com (10 June 2026).

Adjerid, I., Peer, E., and Acquisti, A. (2018). Beyond The Privacy Paradox: Objective Versus Relative Risk in Privacy Decision Making. *MIS quarterly*, 42(2):465–488.

Al Breiki, H. and Mahmoud, Q. H. (2025). A Framework for Integrating Privacy by Design into Generative AI Applications. In *Proceedings of the AAAI Symposium Series*, volume 6, pages 2–9.

Altman, I. (1975). The environment and social behavior: Privacy, personal space, territory, and crowding.

Ambience Healthcare (2026). Ambience Healthcare. Available from https://www.ambiencehealthcare.com (10 June 2026).

Asgarinia, H. (2023). Design for Embedding the Value of Privacy in Personal Information Management Systems. *Journal of Ethics and Emerging Technologies*, 33(1):1–19.

Awan, S. K., Dunoyer, E. J., Genuario, K. E., Levy, A. C., O'Connor, K. P., Serhatli, S., and Gerling, G. J. (2018). Using voice recognition enabled smartwatches to improve nurse documentation. In *2018 Systems and Information Engineering Design Symposium (SIEDS)*, pages 159–164.

Ayalon, O. and Toch, E. (2021). User-centered privacy-by-design: Evaluating the appropriateness of design prototypes. *International Journal of Human-Computer Studies*, 154:102641.

Bain, M., Huh, J., Han, T., and Zisserman, A. (2023). WhisperX: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Balboni, P. and Macenaite, M. (2013). Privacy by design and anonymisation techniques in action: Case study of Ma3tch technology. *Computer law & security review*, 29(4):330–340.

Ball, K. (2010). Workplace surveillance: An overview. *Labor History*, 51(1):87–106.

Ball, K. (2021). Electronic monitoring and surveillance in the workplace: Literature review and policy recommendations.

Ball, K. S. and Margulis, S. T. (2011). Electronic monitoring and surveillance in call centres: A framework for investigation. *New technology, work and employment*, 26(2):113–126.

Bartoletti, I. (2019). AI in healthcare: Ethical and privacy challenges. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 7–10. Springer.

Boaz, A., Hanney, S., Borst, R., O'Shea, A., and Kok, M. (2018). How to engage stakeholders in research: Design principles to support improvement. *Health research policy and systems*, 16(1):60.

Brandeis, L. and Warren, S. (1890). The right to privacy. *Harvard law review*, 4(5):193–220.

Bredin, H. (2023). Pyannote.audio 2.1 speaker diarization pipeline: Principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., and Askell, A. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Bruff, O. and Groves, L. (2026). Scribe and prejudice: Exploring the use of AI transcription tools in social care.

Burgoon, J. K. (1982). Privacy and communication. *Annals of the International Communication Association*, 6(1):206–249.

Cavoukian, A. (2010). Privacy by design: The definitive workshop. A foreword by Ann Cavoukian, Ph. D. *Identity in the Information Society*, 3(2):247–251.

Chen, K., Zhou, X., Lin, Y., Feng, S., Shen, L., and Wu, P. (2025). A survey on privacy risks and protection in large language models. *Journal of King Saud University Computer and Information Sciences*, 37(7):163.

Contissa, G., Lagioia, F., and Sartor, G. (2017). The Ethical Knob: Ethically-customisable automated vehicles and the law. *Artificial Intelligence and Law*, 25(3):365–378.

Corbeil, J.-P., Abacha, A. B., Michalopoulos, G., Swazinna, P., Del-Agua, M., Tremblay, J., Daniel, A. J., Bader, C., Cho, K., and Krishnan, P. (2025). Empowering healthcare practitioners with language models: Structuring speech transcripts in two real-world clinical applications. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 859–870.

Danezis, G., Domingo-Ferrer, J., Hansen, M., Hoepman, J.-H., Metayer, D. L., Tirtea, R., and Schiffner, S. (2015). Privacy and data protection by design-from policy to engineering. *arXiv preprint arXiv:1501.03726*.

Dhagarra, D., Goswami, M., and Kumar, G. (2020). Impact of trust and privacy concerns on technology acceptance in healthcare: An Indian perspective. *International journal of medical informatics*, 141:104164.

Dinev, T. and Hart, P. (2006). An extended privacy calculus model for e-commerce transactions. *Information systems research*, 17(1):61–80.

Distler, V., Lallemand, C., and Koenig, V. (2020). How acceptable is this? How user experience factors can broaden our understanding of the acceptance of privacy trade-offs. *Computers in Human Behavior*, 106:106227.

Draper, N. A. and Turow, J. (2019). The corporate cultivation of digital resignation. *New media & society*, 21(8):1824–1839.

Ehrler, F., Wu, D. T. Y., Ducloux, P., and Blondon, K. (2021). A mobile application to support bedside nurse documentation and care: A time and motion study. *JAMIA Open*, 4(3):ooab046.

Fleischer, S., Berg, A., Zimmermann, M., Wüste, K., and Behrens, J. (2009). Nurse-patient interaction and communication: A systematic literature review. *Journal of Public Health*, 17(5):339–353.

Friedman, B., Kahn, P., and Borning, A. (2002). Value sensitive design: Theory and methods. *University of Washington technical report*, 2(8):1–8.

Friedman, B., Kahn Jr, P. H., Borning, A., and Huldtgren, A. (2013). Value sensitive design and information systems. In *Early Engagement and New Technologies: Opening up the Laboratory*, pages 55–95. Springer.

Gemma Team (2025). Gemma 3 Technical Report.

Ghatnekar, S., Faletsky, A., and Nambudiri, V. E. (2021). Digital scribe utility and barriers to implementation in clinical practice: A scoping review. *Health and Technology*, 11(4):803–809.

Grabler, R., Starzinger, M., Hirschmanner, M., and Frijns, H. A. (2025). Privacy-sensitive conversations between care workers and care home residents in a residential care home (Dataset, Version 1.0.0).

Grillaert, K., Vocke, E., Scarpino, J., Chung, E. Y., and Winston, T. (2025). Applying Privacy-Enhancing Technologies to LLMs in Critical Infrastructure Contexts. In *2025 Cyber Awareness and Research Symposium (CARS)*, pages 1–5. IEEE.

Guo, Y., Wang, J., Hu, D., Tam, S., Gilman, C., Chow, E., Perret, D., Pandita, D., and Zheng, K. (2026). Evaluating ambient artificial intelligence documentation: Effects on work efficiency, documentation burden, and patient-centered care. *Journal of the American Medical Informatics Association*, 33(2):273–282.

Gürses, S. (2014). Can you engineer privacy? *Communications of the ACM*, 57(8):20–23.

Hamdhana, D., Harada, K., Oshita, H., Sakashita, S., and Inoue, S. (2024). Toward extracting care records from transcriptions of visiting nurses using large language models. *International Journal of Activity and Behavior Computing*, 2024(2):1–17.

Hoffmann, C. P., Lutz, C., and Ranzini, G. (2016). Privacy cynicism: A new approach to the privacy paradox. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 10(4).

Höglander, J., Holmström, I. K., Lövenmark, A., Van Dulmen, S., Eide, H., and Sundler, A. J. (2023). Registered nurse–patient communication research: An integrative review for future directions in nursing research. *Journal of Advanced Nursing*, 79(2):539–562.

Hsieh, H.-F. and Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative health research*, 15(9):1277–1288.

Hughes, J., King, V., Rodden, T., and Andersen, H. (1994). Moving out from the control room: Ethnography in system design. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, pages 429–439.

Hutchinson, H., Mackay, W., Westerlund, B., Bederson, B. B., Druin, A., Plaisant, C., Beaudouin-Lafon, M., Conversy, S., Evans, H., and Hansen, H. (2003). Technology probes: Inspiring design for and with families. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 17–24.

Johannesen, L. M. and Helleso, R. (2019). Native language nurses and second language nurses document information in patient records differently. *Sykepleien Forskning*, 14:76342.

Johnson, L. G., Macieira, T. G., Madandola, O. O., Priola, K. J., and Keenan, G. M. (2025). Charting the path forward: Nursing perspectives on documentation and change. *Nursing Outlook*, 73(4):102463.

Jonnagaddala, J. and Wong, Z. S.-Y. (2025). Privacy preserving strategies for electronic health records in the era of large language models. *npj Digital Medicine*, 8(1):34.

Jordan, B. (1996). Ethnographic workplace studies and CSCW. In *Human Factors in Information Technology*, volume 12, pages 17–42. Elsevier.

Kaisler, R. E. and Missbach, B. (2020). Co-creating a patient and public involvement and engagement 'how to'guide for researchers. *Research involvement and engagement*, 6(1):32.

Kanaparthy, N. S., Villuendas-Rey, Y., Bakare, T., Diao, Z., Iscoe, M., Loza, A., Wright, D., Safranek, C., Faustino, I. V., Brackett, A., Melnick, E. R., and Taylor, R. A. (2025). Real-World Evidence Synthesis of Digital Scribes Using Ambient Listening and Generative Artificial Intelligence for Clinician Documentation Workflows: Rapid Review. *JMIR AI*, 4:e76743–e76743.

Keenan, G. M., Yakel, E., Tschannen, D., and Mandeville, M. (2008). Documentation and the nurse care planning process. In *Patient Safety and Quality: An Evidence-Based Handbook for Nurses*. Agency for Healthcare Research and Quality (US).

Kletter, R., Hirschmanner, M., Helena Anna, F., Evelyn, M.-H., and Koeszegi, S. T. (2026). Co-Design and Creation of a Documentation Support System for Care Workers with Large Language Models. *PREPRINT (Version 1) available at Research Square*.

Kletter, R. and Koeszegi, S. T. (2026a). Privacy at the Core: Toward Automated Detection of Privacy-Sensitive Content. *Intelligent Human Systems Integration: Disruptive and Innovative Technologies*, page 84.

Kletter, R. and Koeszegi, S. T. (2026b). Rethinking, Not Reengineering: The Challenge of Digital Documentation in Care. In *Proceedings of the 24th EUSSET Conference on Computer-Supported Cooperative Work (ECSCW) – Conference Papers*. Reports of the European Society for Socially Embedded Technologies.

Kokolakis, S. (2017). Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. *Computers & security*, 64:122–134.

Kumar, D., Choudhary, A., and Kumar, A. (2025). Ethics by Design: Integrating Moral Frameworks in the Digital Age. In *2025 IEEE International Conference on Contemporary Computing and Communications (InC4)*, pages 1–6. IEEE.

Kwak, A., Morrison, C., Bambauer, D., and Surdeanu, M. (2024). Classify first, and then extract: Prompt chaining technique for information extraction. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 303–317.

Kwame, A. and Petrucka, P. M. (2021). A literature-based study of patient-centered care and communication in nurse-patient interactions: Barriers, facilitators, and the way forward. *BMC Nursing*, 20(1):158.

Liu, S.-F., Chen, B.-L., Ding, Y.-L., Yeh, Y.-J., Wang, M.-S., and Su, P.-C. (2025). Application of Automatic Speech Recognition Systems in Telehealth Centers. In *2025 Second International Conference on Artificial Intelligence for Medicine, Health and Care (AIxMHC)*, pages 166–167. IEEE.

Macdonald, L. M. (2016). Expertise in everyday nurse–patient conversations: The importance of small talk. *Global qualitative nursing research*, 3:2333393616643201.

Microsoft (2026). Microsoft Dax Copilot. Available from https://support.microsoft.com/de-de/dax-copilot/learn-dax-copilot-for-dragon-medical-one (10 June 2026).

Ng, J. J. W., Wang, E., Zhou, X., Zhou, K. X., Goh, C. X. L., Sim, G. Z. N., Tan, H. K., Goh, S. S. N., and Ng, Q. X. (2025). Evaluating the performance of artificial intelligence-based speech recognition for clinical documentation: A systematic review. *BMC Medical Informatics and Decision Making*, 25(1):236.

Nissenbaum, H. (2004). Privacy as contextual integrity. *Wash. L. Rev.*, 79:119.

Okon, S. U., Olateju, O., Ogungbemi, O. S., Joseph, S., Olisa, A. O., and Olaniyi, O. O. (2024). Incorporating privacy by design principles in the modification of AI systems in preventing breaches across multiple environments, including public cloud, private cloud, and on-prem. *Including Public Cloud, Private Cloud, and On-prem (September 03, 2024)*.

Ollama (2025). Ollama model registry: Gemma3:27b-it-qat. Available from https://ollama.com/library/gemma3:27b-it-qat (27 July 2026).

Petterson, A., Jaimes Rodriguez, I., Doggett, O., and Chandra, P. (2024). Networks of care in digital domestic labour economies. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–16, Honolulu HI USA. ACM.

Plaquet, A. and Bredin, H. (2023). Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*.

Randall, D. and Rouncefield, M. (2018). Ethnographic approach to design. *The wiley handbook of human computer interaction*, 1:125–141.

Raynes-Goldie, K. (2010). Aliases, creeping, and wall cleaning: Understanding privacy in the age of Facebook. *First Monday*, 15(1).

Reyes-Cruz, G., Spors, V., Muller, M., Ciolfi Felice, M., Bardzell, S., Williams, R. M., Hansson, K., and Feldfeber, I. (2025). Resisting AI Solutionism: Where Do We Go From Here? In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–6.

Rotenstein, L., Melnick, E. R., Iannaccone, C., Zhang, J., Mugal, A., Lipsitz, S. R., Healey, M. J., Holland, C., Snyder, R., Sinsky, C. A., Ting, D., and Bates, D. W. (2024). Virtual Scribes and Physician Time Spent on Electronic Health Records. *JAMA Network Open*, 7(5):e2413140.

Sadek, M., Constantinides, M., Quercia, D., and Mougenot, C. (2024). Guidelines for integrating value sensitive design in responsible AI toolkits. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–20.

Seah, C., Ng, T. P., Teo, S. Y., Lee, J.-J., Wu, Y., Zheng, C., Foong, P. S., and Cho, J. (2026). Who Takes Care of Technology?: The Invisible Labor of Infrastructuring Digital Technologies in Aged Care. *Proceedings of the ACM on Human-Computer Interaction*, 10(1):1–39.

Shattell, M. (2004). Nurse–patient interaction: A review of the literature. *Journal of Clinical Nursing*, 13(6):714–722.

Sun, S., Yuan, R., Cao, Z., Li, W., and Liu, P. (2024). Prompt chaining or stepwise prompt? refinement in text summarization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7551–7558.

Sundler, A. J., Hjertberg, F., Keri, H., and Holmström, I. K. (2020). Attributes of person-centred communication: A qualitative exploration of communication with older persons in home health care. *International Journal of Older People Nursing*, 15(1):e12284.

Tahera, M. A., Sidhu, K. S., Dass, S., and Saha, S. (2026). SoK: Privacy-aware LLM in Healthcare: Threat Model, Privacy Techniques, Challenges and Recommendations. *arXiv preprint arXiv:2601.10004*.

Tufford, F., Lowndes, R., Struthers, J., and Chivers, S. (2018). 'Call security': Locks, risk, privacy and autonomy in long-term residential care. *Ageing International*, 43(1):34–52.

Ullah, I., Hassan, N., Gill, S. S., Suleiman, B., Ahanger, T. A., Shah, Z., Qadir, J., and Kanhere, S. S. (2024). Privacy preserving large language models: Chatgpt case study based vision and framework. *IET Blockchain*, 4:706–724.

Umbrello, S. and Van de Poel, I. (2021). Mapping value sensitive design onto AI for social good principles. *AI and Ethics*, 1(3):283–296.

Van Buchem, M. M., Boosman, H., Bauer, M. P., Kant, I. M. J., Cammel, S. A., and Steyerberg, E. W. (2021). The digital scribe in clinical practice: A scoping review and research agenda. *npj Digital Medicine*, 4(1):57.

van Rest, J., Boonstra, D., Everts, M., van Rijn, M., and van Paassen, R. (2012). Designing privacy-by-design. In *Annual Privacy Forum*, pages 55–72. Springer.

Van Veen, D., Van Uden, C., Blankemeier, L., Delbrouck, J.-B., Aali, A., Bluethgen, C., Pareek, A., Polacin, M., Reis, E. P., Seehofnerová, A., Rohatgi, N., Hosamani, P., Collins, W., Ahuja, N., Langlotz, C. P., Hom, J., Gatidis, S., Pauly, J., and Chaudhari, A. S. (2024). Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30(4):1134–1142.

Vithanage, D., Yu, P., Xie, Q., Xu, H., Wang, L., and Deng, C. (2025). A comprehensive evaluation of large language models for information extraction from unstructured electronic health records in residential aged care. *Computers in Biology and Medicine*, 197:111013.

Voize (2026). Voize – AI companion for nurses. Available from https://www.voize.ai (10 June 2026).

Wacnik, P., Daly, S. R., and Verma, A. (2025). Participatory design: A systematic review and insights for future practice. *Design Science*, 11:e21.

Wang, Z., Guo, Y., Bowman, E., Zhai, Y., Shu, X., Zhang, S., Helms, K., Capel, T., and Vines, J. (2026). Caring about Care: A Meta-Narrative Review of HCI Research on Care. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–31, Barcelona Spain. ACM.

Zhang, X., Xu, H., Ba, Z., Wang, Z., Hong, Y., Liu, J., Qin, Z., and Ren, K. (2024). Privacyasst: Safeguarding user privacy in tool-using large language model agents. *IEEE Transactions on*

Dependable and Secure Computing, 21(6):5242–5258.

Zhang, Z., Rossi, R. A., Kveton, B., Shao, Y., Yang, D., Zamani, H., Dernoncourt, F., Barrow, J., Yu, T., Kim, S., Zhang, R., Gu, J., Derr, T., Chen, H., Wu, J., Chen, X., Wang, Z., Mitra, S., Lipka, N., Ahmed, N., and Wang, Y. (2025). Personalization of large language models: A survey.

Zimmerman, J., Forlizzi, J., and Evenson, S. (2007). Research through design as a method for interaction design research in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 493–502.

A. LLM-BASED AUTOMATED PRIVACY DETECTION PIPELINE

As with the LDSS, we restrict the privacy detection pipeline to locally deployed LLMs using the Ollama platform (Ollama 2025) to safeguard the processed data. We chose the *gemma3:27b-it-qat* model (Gemma Team 2025) because it offers strong performance while remaining within the limited computational capacity of our LLM workstation. To delegate structured tasks to the LLM we want it to be deterministic (*temperature=0.1*), select only from high-probability tokens (*top_p=0.9*), avoid low-probability tokens to stabilise output (*top_k=40*) and to avoid token repetition (*repeat_penalty=1.05*). To be able to capture the full system and user prompts along side the transcripts we opted for a large 32k context window.

A.1. Data Preparation

The transcripts used in the longitudinal and previous studies were generated by automatically transcribing the audio recordings with *WhisperX* (Bain et al. 2023) which uses *faster-whisper-large-v3* (see Kletter et al. 2026). Automatic speaker diarisation with *pyannote-audio 3* (Plaquet and Bredin 2023; Bredin 2023) did not yield reliable results for this use case. However, the 23 transcripts from the longitudinal deployment include speaker diarisation, indicated by speaker tags (e.g. *SPEAKER_00*, *SPEAKER_01*) at the beginning of each line. In all other transcripts, no such distinction is made.

A.2. Few-Shot Prompting

To compile a list of privacy-sensitive transcript excerpts, the privacy detection pipeline exploits LLMs' ability to excel as few-shot learners (Brown et al. 2020). We inject two to three examples from a publicly accessible database (Grabler et al. 2025) containing synthetic, privacy-sensitive conversations between care workers and care recipients into the prompt. The examples contain a conversation chunk, an assigned privacy category, and the individual whose privacy is affected by the conversation. They were translated locally from English into German, the language of the transcripts, by prompting the *gemma3:27b-it-qat* model to act as a translator. As our transcripts are derived from the residential long-term care context as well, we adopt the same privacy categories as those used in the example database (Grabler et al. 2025):

- *Personal information:* Basic identifying or biographical information about a person.
- *Family:* Information about family members or family relationships.

- *Health:* Information related to physical or mental health.
- *Thoughts:* Private opinions, beliefs, or internal reflections.
- *Values:* Moral principles or ideological beliefs.
- *Acquaintance:* Information about people someone knows or social contacts.
- *Appointment:* Information about scheduled meetings or commitments.

These categories, along with descriptive guidance for assigning them, are injected as an ontology in the system prompt (see Section A.4.1). The user prompt includes the transcript and strict instructions for the expected JSON output (see Section A.4.2). Additionally, the LLM is prompted to provide a confidence score, a reason why the information is privacy-sensitive, and a guess of who is disclosing the information. The minimum candidates was set to 5 and the maximum to 15. The affected individuals in the provided examples are either the *care worker*, the *care recipient* (the resident), *both* when the privacy of both the care worker and the care recipient is affected, or *other* when the privacy of another individual not directly part of the care relationship is affected.

A.3. Candidate-Restricted Selection

From the first iteration, we obtain a set of candidates, including potentially privacy-sensitive conversation segments. Using prompt chaining (Kwak et al. 2024; Sun et al. 2024), we restrict the LLM to selecting only from these pre-selected candidates in the final selection step. We provide only a JSON representation of the pre-selected candidates, which includes the initial reason for inclusion and the predicted speaker, privacy category and affected individual, alongside the unchanged system prompt (see Section A.4.1). The user prompt restricts selection to the provided candidates, requiring an ID reference to the originals (see Section A.4.3). The maximum number of entries was set to 20, to allow for all pre-selected candidates to be included. Additionally, the LLM must indicate whether it agrees with the initial assessment of the candidate being privacy-sensitive. It must include a reason for an affirmative or negative assessment, and again a confidence score. The selection mechanism is applied iteratively until all pre-selected candidates have been assessed, yielding the final set of privacy-sensitive segments.

A.4. Prompt Engineering

The system and user prompts were implemented using Jinja templates. The prompts shown in this paper are English translations of the original German versions.

A.4.1. System Prompt

{*# system_prompt.jinja #*}
You are a precise information extraction assistant.

Your task is to identify all passages in a German–language transcript of a conversation between a care worker and a resident that contain information relevant to privacy.

The speaker identification in the transcript is unreliable, so you'll need to determine who is speaking based on the content of the text. Most of the time, it's the care worker speaking.

Strict rules:
– NO Hallucinations: Extract only information that can be directly verified in the transcript.
– Interpretations based on context are allowed.
– Use categories from the ontology and expand them as needed.
– Always choose the most specific category that applies.

Ontology for classifying privacy categories:

{{ ontology }}

Examples of texts, their categories, and the individuals whose privacy is affected. CW (Care Worker) refers to the care worker, and CR (Care Recipient) refers to the resident:

{% if examples %}
{{ examples }}
{% endif %}

A.4.2. Evidence Selection

{*# evidence_select.jinja #*}
Transcript (with line numbers, in the format "Ln X: ..."):
{{ transcript }}

Task:
A) Select **{{ min_candidates | default(12) }}** to **{{ max_candidates | default(30) }}** text segments from the transcript.
B) Every span:
  – continuous
  – typically 2–8 lines
C) Give preference to content from the classes defined in the ontology.
D) Output direct quotes as verbatim_text.
E) Briefly explain why the range contains information relevant to privacy (reason).

Strict output: ONLY valid JSON, no Markdown, no additional keys.
{% raw %}
```
{
  "candidates": [
    {
      "id": "E1",
      "confidence": 1.0,
      "category": "...",
```

```
        "affected_speaker": "...",
        "speaker": "...",
        "reason": "...",
        "start_line": 1,
        "end_line": 4,
        "verbatim_text": "Ln 1: ...\nLn 2: ..."
      }
  ]
}
{% endraw %}

Definitions:
– category: What kind of privacy–related information is
shared
– speaker: Who is providing the information ('Care_worker', '
Resident', 'Unclear')
– affected_speaker: To whom the information refers ('
Care_worker', 'Resident', 'Third_person', 'Unclear')
```

A.4.3. Candidate-Restricted Selection

```
{# grounded_draft.jinja #}
Text segments (you may only select from these):
{{ candidates_json }}

Task:
Create a maximum of {{ max_entries | default(10) }}
entries.
– Each entry MUST reference 1 candidate_id.
– Do not use any information other than what is in the
supporting text segments.
– Briefly explain why the entries are relevant to privacy (
reason).
– If supported=false, explain why the entries cannot be
supported.

Strict output: ONLY valid JSON.
{% raw %}
{
  "grounded_draft": [
    {
      "candidate_id": "C3",
      "supported": true,
      "confidence": 0.0,
      "reason_if_supported": ""
      "reason_if_not_supported": "",
    }
  ]
}
{% endraw %}

Rules:
– supported=false –> reason_if_not_supported not empty,
reason_if_supported=""
– supported=true –> reason_if_supported not empty,
reason_if_not_supported=""
```