Human Systems Engineering and Design (IHSED 2026), Vol. 237, 2026, 262–272
https://doi.org/10.54941/ahfe1008166

From Generic to Personalized: Lessons for Meaningful Human–AI Interaction in LLM-Supported Care Documentation

Reinhard Kletter and Sabine Theresia Koeszegi

Institute of Management Science, TU Wien, 1040 Vienna, Austria

ABSTRACT

We present an LLM-based documentation support system that extracts care-relevant information from nurse–resident interactions in residential care. While care workers generally responded positively, the inclusion of care worker–related information in reports remains a key issue. We investigate whether personalization via user profile injection can reduce such unwanted output. To this end, we introduce an LLM-based pipeline for automatically extracting care worker facts from transcripts. Results show that both profile injection and candidate-constrained prompting reduce care worker information, but lead to lower evaluation scores, indicating limited post-editing of the initial reports. We therefore argue for incorporating reflection mechanisms to mitigate automation bias, as well as feedback loops to further improve system performance and enable adaptation to individual documentation preferences.

Keywords: LLMs, Care work, Documentation support, Personalization, Human-AI interaction

INTRODUCTION

Digitalization has increasingly permeated the nursing profession. While electronic health records (EHRs) offer clear benefits (Kruse et al., 2017), many care workers still perceive their use as burdensome (Johnson et al., 2025). To alleviate some of this burden, we developed an LLM-based documentation support system (LDSS) that extracts care-relevant information from audio recordings of nurse–resident interactions. We adapted our previous work (Kletter et al., 2026) and deployed the system in a residential care home for a total of 22 weeks to investigate real-world usage patterns and gather user feedback. Our findings indicate that the primary issue affecting report quality stems from unreliable automatic speaker diarization in this use case: care workers are often overrepresented because the recording device is attached to their clothing and captures their speech more clearly, while residents tend to speak more softly and background noise further complicates speaker distinction. This causes the LDSS to include care worker–related information in the generated reports, reducing its utility by increasing the required post-editing effort.

To address this, we investigate alternative approaches based on LLM personalization (Zhang et al., 2025). Specifically, we inject a care worker–specific user profile to support the LDSS in distinguishing between speakers.

Received April 12, 2026 Revised June 24, 2026 Accepted July 8, 2026 Available online September 25, 2026

We introduce the LLM-based user profile development (LUPD) pipeline, which automatically extracts care worker–related information from prior transcripts to construct such profiles. In addition, we evaluate different prompting strategies.

Our research question is: *"How can LLM-supported care documentation be technically improved to promote meaningful human–AI interaction?"* In this work, we present a technical framework for extracting care-relevant information from nurse–resident interactions in a residential nursing home. We propose an approach for automatically generating user profiles from interaction transcripts and evaluate whether personalized prompting improves system performance. Building on this, we outline feedback mechanisms for the continuous refinement of these profiles, enabling adaptation to individual documentation preferences. Finally, based on observed usage patterns, we argue that reflection mechanisms are essential for fostering effective human–AI collaboration in documentation tasks.

RELATED WORK

Advances in automatic speech recognition (ASR) and large language models (LLMs) have enabled promising applications for documentation support in the healthcare sector. However, most existing research has focused on supporting physicians (Du et al., 2024; Rotenstein et al., 2024), with comparatively little attention given to nurse–patient interactions. These interactions are typically longer and often embed clinically relevant information within informal conversation or small talk (Macdonald, 2016). Recent work by Hamdhana et al. (2024) leverages ASR and LLMs to extract care-relevant information from online visits, while Liu et al. (2025) focus on telehealth interventions. Other approaches explore the use of LLMs to structure nurse dictations (Corbeil et al., 2025) or to organize EHR information in residential care settings (Vithanage et al., 2025).

METHODOLOGY

LLM-Based Documentation Support System

Research Approach. We followed co-design and value sensitive design (VSD) approaches to ensure that the system is meaningful, usable, and aligned with human–centered values. We involved stakeholders at ideation, development and testing, and iteratively implemented their feedback.

Tech Setup. The LDSS consists of a recording unit, an upload unit, and an LLM workstation for processing audio and transcripts. Care workers carry the recording unit—a device with local storage—during care activities. The upload unit, a Raspberry Pi–based system with a 10.1-inch touchscreen, is supplemented with an external mouse and keyboard upon care worker request. During upload, care workers assign each recording to a resident and select a mode: *interaction* or *dictation*. The files are then securely transmitted to an LLM workstation at the researchers' university. Once uploaded, audio files are automatically deleted from the recording unit for

privacy reasons. Audio is transcribed locally using *WhisperX* (with *faster-whisper-large-v3*) and diarized with *pyannote-audio 3*. Structured, bullet-point reports are generated using the *gemma3:27b-it-qat* model with zero-shot prompting, with temperature set to 0.1 to reduce output variability. Reports are made available in the care organization's cloud as editable Word files. Care workers receive both the generated report as reference and an editable version for transfer to the documentation program. Additionally, they complete evaluations, including 5-point Likert-scale ratings of quality and comparison to their own reports, the usable share (0–100%), and optional free-text feedback (for a more detailed description, see Kletter and Koeszegi, 2026b).

Modes. The LDSS supports multiple modes: (i) *Interaction* (recorded during care) or *dictation* (recorded afterwards): Interaction transcripts include speaker tags (e.g., *SPEAKER_00*) from automatic diarization, and prompts are adapted accordingly to guide summarization. (ii) *Merge transcripts* vs. *merge results*: By default, recordings for a resident are merged per day to generate a daily report. For weekly reports, summaries from the past seven days are combined, prioritizing care worker–edited versions over system outputs.

Prompting. The user prompt encloses the transcript between tags and requests a care documentation summary, with constraints: seven key points (ending with "additional observations"), passive phrasing from the care worker's perspective, use of common abbreviations, focus on the resident, and no verbose introductions. The system prompt defines the model as an assistant for generating care documentation from German transcripts and explains mode-specific tags. It also includes instructions—derived from our prior work (Kletter et al., 2026)—on what constitutes care-relevant information, what is recorded elsewhere (e.g., the checkbox-based *proof of implementation*), and what to exclude.

LLM-Based User Profile Development

Data. Across testing, evaluation, and longitudinal deployment of the LDSS, 60 transcripts (≈200 lines each) were collected in one residential nursing home, involving eight care workers and ten residents. We grouped the data into four clusters: transcripts for developing the initial prototype (C1); transcripts evaluated with care workers, who annotated usefulness and relevance (C2); transcripts from longitudinal deployment without feedback (C3), and with care worker feedback and edited reports for documentation (C4). Clusters 3 and 4 include automated diarization; transcripts were automatically merged by day, though there are rarely multiple recordings by day. User profile creation focuses on one care worker (CW1) who contributed to both C2 and C4. The profile is built from 13 transcripts (C1: 1, C2: 6, C3: 6) and evaluated on the 8 edited reports from C4 using a consistent prompt setup (see Table 1).

LUPD Pipeline. We build on care worker annotations from Cluster 2 but simplify labeling by assigning tags at the bullet level rather than to text spans. Using these annotated examples, we extract care worker–related

information from transcripts in Clusters 1–3, excluding Cluster 4 for evaluation of profile injection. The LUPD pipeline comprises five steps, leveraging prompting techniques such as in-context examples (Brown et al., 2020) and prompt chaining (Kwak et al., 2024) to improve accuracy: (1) identify candidates containing care worker information using in-context examples and an ontology of categories (e.g., personal data, events) with patterns; (2) draft bullets from candidates; (3) verify them against the source to prevent hallucinations; (4) repair and re-verify unsupported bullets; (5) reformulate validated bullets into care worker facts and structure based on a target ontology. We use a near-deterministic, stabilized LLM setup (temperature = 0.1, top_p = 0.9, top_k = 40) with a repeat penalty of 1.05 to reduce repetition.

Evaluation Metrics

To compare LDSS-generated reports with care worker–edited versions, we use Translation Edit Rate (TER) for post-editing effort, ROUGE-1 for lexical overlap, and BERTScore (F1) for semantic similarity. Personalization effects are evaluated by injecting LUPD-derived user profiles into the original prompts for Cluster 4 transcripts, while keeping the LLM settings unchanged. We also assess candidate-constrained prompting—where candidate selection precedes bullet generation—both with and without profile injection, using the same near-deterministic, stabilized, and non-repetitive LLM settings as in the LUPD. Additionally, we annotate care worker information in the outputs as *mention* or *conflation* (i.e., mixed with resident information) to analyze relative changes across setups.

RESULTS

Table 1 summarizes both the care workers' (CW) evaluations of the system, and the computed scoring metrics comparing generated reports to care worker–corrected versions. All evaluations are based on daily reports. In the subjective ratings (quality, comparison, and usable share), CW1 assigned slightly higher scores than CW2; however, somewhat contradictorily, CW1's calculated scores are worse on average. Overall, the written feedback was positive, report quality was rated above average, the generated reports were perceived as an improvement over those produced without the system, and, on average, approximately three quarters of the generated content was considered usable. E1–E3 used a preliminary prompt setup that did not yet include explicit instructions to distinguish care-relevant from non-relevant information. The report of E4 was generated without an underlying transcript, as the workstation ran out of memory during the ASR process. Although the resulting report contained only generic statements (e.g., no complaints, pain, behavioral or mood changes, unmet needs, or changes in independence), it was rated highly by CW1 and received notably positive written feedback. For evaluations E2, E3, E6, E7, E8, and E10, CW1 reported instances in which their own information was either explicitly mentioned or conflated with that of the resident.

Table 1: Care worker evaluations (CW1, CW2) and computed metrics of generated reports (Cluster 4) for residents (R). (∗) modified prompt setup. (†) generated without transcript. *Note.* Parts of this table were previously published in Kletter and Koeszegi (2026b).

ID	Care Worker	Resident	Quality	Comparison	Usable Share	TER	ROUGE-1	BERT Score
E1*	CW1	R1	4.00	3.00	80%	0.1974	0.8889	0.8071
E2*	CW1	R1	4.00	3.00	80%	0.1299	0.9427	0.8615
E3*	CW1	R1	3.00	3.00	70%	0.8836	0.3178	0.1153
E4†	CW1	R1	4.00	4.00	90%	0.1905	0.9155	0.9020
E5	CW1	R1	4.00	4.00	80%	0.1959	0.8716	0.8365
E6	CW1	R1	4.00	4.00	70%	1.2295	0.5464	0.5730
E7	CW1	R1	4.00	4.00	100%	1.1781	0.5893	0.5406
E8	CW1	R1	4.00	4.00	100%	0.4842	0.7742	0.6910
E9	CW1	R2	5.00	4.00	100%	0.1712	0.8767	0.8256
E10	CW1	R1	2.00	2.00	30%	1.6885	0.3109	0.1466
E11	CW1	R1	4.00	4.00	70%	0.5149	0.6140	0.5141
E12	CW1	R2	5.00	5.00	90%	0.1692	0.8864	0.8113
Σ12	**Mean**		**3.92**	**3.67**	**80%**	**0.5861**	**0.7112**	**0.6354**
	±SD		**± 0.79**	**±0.78**	**±20%**	**±0.5307**	**±0.2298**	**±0.2696**
E13	CW2	R3	4.00	3.00	80%	0.0667	0.9424	0.9055
E14	CW2	R4	2.00	4.00	40%	0.6667	0.5472	0.5045
E15	CW2	R5	3.00	2.00	40%	0.0933	0.9259	0.9054
E16	CW2	R6	4.00	2.00	100%	0.1802	0.8512	0.8444
E17	CW2	R5	4.00	4.00	80%	0.3196	0.8658	0.7944
Σ5	**Mean**		**3.40**	**3.00**	**68%**	**0.2653**	**0.8265**	**0.7909**
	± SD		**±0.89**	**±1.00**	**±27%**	**±0.2451**	**±0.1609**	**±0.1667**
Σ17	**Mean**		**3.76**	**3.47**	**76%**	**0.4917**	**0.7451**	**0.6811**
(All)	**± SD**		**±0.83**	**±0.87**	**±22%**	**±0.4810**	**±0.2138**	**±0.2495**

Figure 1 shows the distribution of the different scores for each care worker individually and combined. For the combined evaluations, the TER score has its median at $M_d = 0.1974$, indicating that for around half of the generated reports only little post-editing effort is required. The ROUGE-1 score indicates a high lexical overlap ($M_d = 0.8658$) for a substantial portion of the reports. Similarly, the BERTScore suggests frequent semantic similarity between generated and reference reports ($M_d = 0.8071$). At the same time, the distributions exhibit considerable spread across all three metrics, and differ across care workers. This variability indicates that system performance depends strongly on characteristics of the underlying transcripts, such as their length, structure, and how clearly information can be attributed to residents. Cases with poor scores typically correspond to instances where care workers needed to adapt or remove care worker–related information, or discard bullets that were deemed irrelevant.

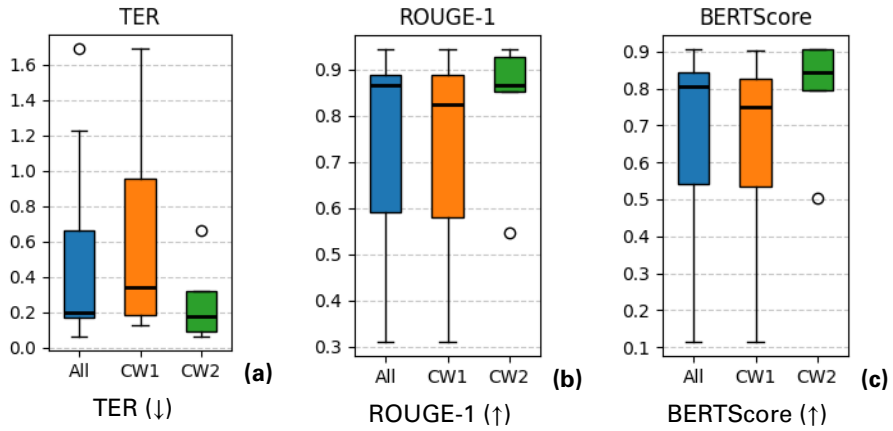

Figure 1: Distribution of scores across all evaluations (E1–E17) for the initial configuration. Lower TER and higher ROUGE-1 and BERTScore indicate better performance.

Profile Injection and Alternative Prompting Techniques

User Profile. The LUPD successfully extracted several aspects of the care worker's personal context. This included information about their family situation (e.g., marital status, presence of children, relationships with parents, and references to nephews or nieces). Additionally, it inferred the existence and type of a pet, as well as certain personal preferences, primarily related to weather. The pipeline also identified indications of a high workload and resulting exhaustion. However, some relevant information was not captured accurately. Notably, no names were extracted despite being present in the transcripts. Furthermore, the system incorrectly assumed a spoken language, while the actual mother tongue could have been inferred from contextual cues but was not. Similarly, the care worker's origin could likely have been inferred, as the respective country is occasionally referenced in both the transcripts and the generated reports.

Score Comparison. Compared to the care worker–corrected reports, the alternative setups—zero-shot with profile (ZS+P), candidate-constrained without (CO), and with profile (CO+P)—yield, on average, worse scores than the original zero-shot setup without profile (ZS) (see Table 2). This trend is also reflected in the score distributions (see Figure 2). Among the evaluated setups, ZS+P yields the closest results, reflecting the similarity in the prompting setup.

Table 2: Average TER, ROUGE-1, and BERTScore and standard deviation for E5–E12.

Prompting Setup	TER	ROUGE-1	BERTScore
Zero-Shot w/o Profile (ZS)	0.7039 ± 0.583	0.6837 ± 0.205	0.6174 ± 0.231
Zero-Shot w/ Profile (ZS+P)	1.0632 ± 0.405	0.4409 ± 0.117	0.2856 ± 0.177
Constrained w/o Profile (CO)	1.1517 ± 0.447	0.3342 ± 0.074	0.1179 ± 0.101
Constrained w/ Profile (CO+P)	1.1824 ± 0.399	0.3315 ± 0.079	0.1018 ± 0.088

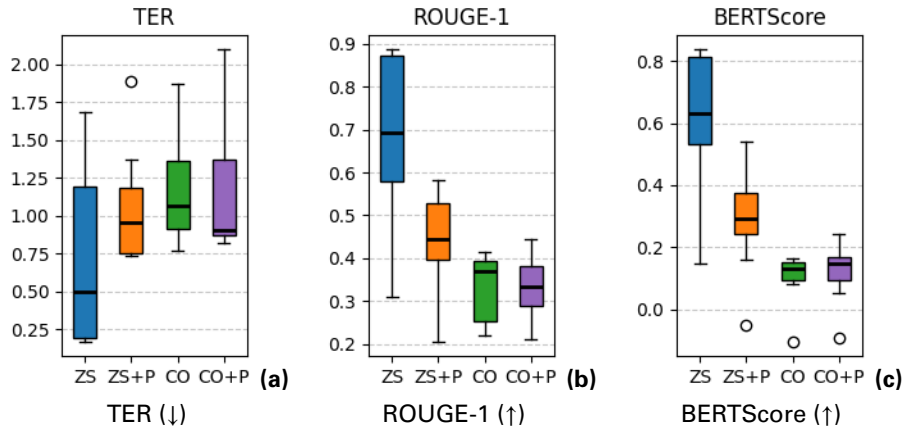

Figure 2: Distribution of TER, ROUGE-1, and BERTScore across 8 evaluations (E5–E12) for all configurations.

Content Comparison. Care worker–related information in the generated reports decreased from 15 instances in the ZS setup to 10 in ZS+P, 11 in CO, and 13 in CO+P. Table 3 further reports the average relative share of such information with respect to the total number of report lines, both overall and disaggregated by whether it was explicitly mentioned or conflated with resident information. Taking the ZS setup as the baseline, this corresponds to an average reduction of approximately 20–33.3% across the alternative configurations. Although most evaluations show a reduction in absolute instances, in E6 the constrained setups contain more care worker–related information than the baseline.

Table 3: Mean and standard deviation of contained care worker information rates in different setups for 8 evaluations (E5–E12). Improvements (in overall) are relative to the ZS baseline.

Prompting Setup	Overall	Mention	Conflation
ZS	26.8 ± 29.8%	5.4 ± 9.9%	21.4 ± 26.7%
ZS+P	17.9 ± 23.4% (↓33.3%)	7.1 ± 12.4%	10.7 ± 18.6%
CO	18.8 ± 24.5% (↓30.0%)	1.6 ± 4.1%	17.2 ± 22.8%
CO+P	21.4 ± 28.3% (↓20.0%)	0.0 ± 0.0%	21.4 ± 28.3%

DISCUSSION

Improving Through Personalization

The results indicate that extracting care-relevant information from care interactions to support written documentation is feasible. However, performance varies both across care workers and across individual recordings or transcripts. Furthermore, Table 1 does not reveal consistent patterns across specific care worker–resident pairings. The primary source of errors—manifesting in increased editing effort, lower semantic similarity, and overall

poorer care worker ratings—stems from the inclusion of care worker–related information in the generated reports. Injecting a constructed user profile into the original zero-shot prompt had the strongest effect in reducing content. In contrast, within the candidate-constrained setups, the version with profile injection performed slightly worse.

Although the observed improvements are modest, even eliminating the need to edit a single bullet point per report represents a meaningful benefit for care workers using the LDSS. Accordingly, incorporating profile injection in a zero-shot setup appears promising. This approach could be further enhanced by introducing feedback mechanisms that enable care workers to easily flag self-referential content and other quality issues, thereby continuously refining the profile and improving subsequent outputs. Over time, this would also support adaptation to individual documentation preferences.

While both candidate-constrained setups reduced the inclusion of care worker–related information, prompt chaining on the local LLMs led to substantially longer generation times, making this approach less practical. Similarly, the LUPD pipeline—also based on multi-step prompt chaining—is time-intensive. To address this, direct editing of the user profile should be enabled for care workers, also allowing them to manage their individual privacy needs.

Human–AI Interaction

While the computed scores do not capture qualitative aspects of the content, a notable pattern emerges: all scores differ substantially from the baseline—that is, from those of the initially generated reports (Figure 2). This suggests that care workers generally made only minor modifications to the generated reports. This observation is further supported by the initial data, where CW2 introduced only limited changes (see Table 1 and Figure 1). However, it cannot be ruled out that, despite the instructions, time constraints prevented care workers from fully revising the reports to a truly final, usable state. At the same time, CW2's combination of minimal edits and lower qualitative ratings suggests a potential inconsistency between editing behavior and evaluation. However, even CW1 who demonstrably invested time in editing still left large portions of the reports very similar to the generated versions, reinforcing the notion that the system output was already close to acceptable in many cases.

LLMs are known to produce highly convincing outputs that may discourage critical scrutiny (Swanson et al., 2026), increasing the risk of automation bias—i.e., users accepting automated suggestions uncritically (Qazi et al., 2025). In this context, generated reports may appear complete, even though important aspects—such as non-verbal cues observable only by care workers—are inherently not captured. This effect is illustrated by the report generated solely from the system and user prompt, without an underlying transcript (see Table 1). Although the resulting structured output contained no concrete information from the interaction, it was still perceived positively by the care worker, as it did not include obvious errors or incorrect statements—despite effectively conveying no meaningful

content. This tendency may be further reinforced by care workers—many of whom are second-language speakers—perceiving their own language skills as inferior to those of the LLM, making it more difficult to revise text that already appears polished. Such self-doubt can contribute to overtrust in AI-generated outputs (Vaccaro et al., 2024). Supporting this interpretation, the corrected reports—while improved in terms of content—often introduced spelling and grammatical errors that were not present in the original generated versions.

Studies have shown that the use of LLMs can entail cognitive costs when not applied appropriately, potentially limiting the formation of deeper neural connectivity patterns (Kosmyna et al., 2025). In this context, it is important to acknowledge that the observed use of the LDSS may—similar to other AI systems (Handa et al., 2025)—encourage only minimal engagement with generated content. This, in turn, may lead care workers to reflect less on care measures and develop a reduced depth of knowledge about residents.

We therefore argue for the integration of mechanisms that actively promote reflection within the LDSS. Drawing on the concept of Reflection Machines (Haselager et al., 2024), care workers should be prompted with occasional, well-designed questions when reviewing generated reports, encouraging them to recall subtle details from interactions and refine the content accordingly. Crucially, such mechanisms should be personalized to maximize their effectiveness (Fügener et al., 2021) and designed to be perceived as supportive rather than intrusive or controlling. These measures enable meaningful human–AI interaction, moving beyond basic collaboration toward true synergy (Vaccaro et al., 2024), rather than replacing the unique knowledge and skills of care workers with overly solutionist AI approaches.

LIMITATIONS

The study is limited by its small sample size. Participation adds workload and some care workers expressed concerns about potential organizational monitoring through audio recordings. Consequently, the findings have limited generalizability to other organizational or national contexts.

Future Work. Creating user profiles raises inherent privacy concerns that must be addressed. In addition to our proposed detection of privacy-sensitive information in transcripts (Kletter and Koeszegi, 2026a), clear access controls and robust security measures are required. Beyond using transcripts as-is and injecting user profiles, future work should explore more reliable methods for attributing speaker roles at the transcript-line level, rather than only during summarization.

CONCLUSION

Our findings show that the proposed LDSS can extract care–relevant information for written reports with acceptable quality. Personalization—via automatically generated user profiles capturing personal attributes—helps reduce care worker–related content. Candidate-constrained prompting achieves similar reductions but is slower than the zero-shot

baseline. Evaluation metrics indicate minimal post-editing by care workers. We therefore recommend incorporating reflection mechanisms to mitigate automation bias and preserve individual expertise, alongside feedback loops to improve output quality and adapt to care workers' documentation practices.

ACKNOWLEDGMENT

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.

REFERENCES

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.

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', pp. 859–870.

Du, X., Wang, Y., Zhou, Z., Chuang, Y.-W., Yang, R., Zhang, W., Wang, X., Zhang, R., Hong, P., Bates, D. W. and Zhou, L. (2024), 'Generative large language models in electronic health records for patient care since 2023: A systematic review'.

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'.

Fügener, A., Grahl, J., Gupta, A. and Ketter, W. (2021), 'Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI', Management Information Systems Quarterly (MISQ)-Vol 45.

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.

Handa, K., Bent, D., Tamkin, A., McCain, M., Durmus, E., Stern, M., Schiraldi, M., Huang, S., Ritchie, S., Syverud, S., Jagadish, K., Vo, M., Bell, M. and Ganguli, D. (2025), 'Anthropic education report: How university students use claude', https://www.anthropic.com/news/anthropic-education-report-howuniversity-students-use-claude.

Haselager, P., Schraffenberger, H., Thill, S., Fischer, S., Lanillos, P., Van De Groes, S. and Van Hooff, M. (2024), 'Reflection machines: Supporting effective human oversight over medical decision support systems', Cambridge Quarterly of Healthcare Ethics 33(3), 380–389.

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.

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

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 p. 84.

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I. and Maes, P. (2025), 'Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task', arXiv preprint arXiv:2506.08872 4.

Kruse, C. S., Mileski, M., Vijaykumar, A. G., Viswanathan, S. V., Suskandla, U. and Chidambaram, Y. (2017), 'Impact of electronic health records on long-term care facilities: Systematic review', JMIR medical informatics 5(3), e35.

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', pp. 303–317.

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)', IEEE, pp. 166–167.

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

Qazi, I. A., Ali, A., Khawaja, A. U., Akhtar, M. J., Sheikh, A. Z. and Alizai, M. H. (2025), 'Automation Bias in Large Language Model Assisted Diagnostic Reasoning Among AI-Trained Physicians', medRxiv pp. 2025–08.

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.

Swanson, K., Bent, D., Ludwig, Z., Dakan, R. and Feller, J. (2026), 'Anthropic education report: The AI fluency index', https://www.anthropic.com/news/anthropic-education-report-the-ai-fluencyindex.

Vaccaro, M., Almaatouq, A. and Malone, T. (2024), 'When combinations of humans and AI are useful: A systematic review and meta-analysis', Nature Human Behaviour 8(12), 2293–2303.

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.

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'.