MEDIVOICE: Next-generation AI for clinical overload assistant

Shreya S. Kadam
Department of Computer Science Engineering
Yashoda Technical Campus, Satara, Maharashtra, India

MR. Vikas Chavan
Department of Computer Science Engineering
Yashoda TechnicalCampus, Satara, Maharashtra, India

Tanuja D. Thorat
Department of Computer Science Engineering
Yashoda Technical Campus, Satara, Maharashtra, India

Sanika B. Kuchekar
Department of Computer Science Engineering
Yashoda Technical Campus, Satara, Maharashtra, India

Sejal S. Raskar
Department of Computer Science Engineering
Yashoda Technical Campus, Satara, Maharashtra, India

Abstract— The increasing workload in healthcare creates challenges in clinical documentation and patient management. This project proposes an AI-Based Clinical Overload Assistant to automate medical documentation using Artificial Intelligence. The system uses Speech-to-Text technology to convert doctor–patient conversations into text in real time. Medical NLP extracts important details such as symptoms, diagnoses, medications, and treatment plans. Based on the extracted data, the system automatically generates draft prescriptions and clinical notes. The assistant also asks important follow-up questions if any medical information is missed during consultation. The system supports multiple global languages, making it useful in diverse healthcare environments. Real-time processing improves consultation efficiency and reduces documentation errors. By minimizing manual work, the system allows doctors to focus more on patient care. Overall, the project provides a smart, scalable, and efficient AI solution for modern healthcare management.

INTRODUCTION

Doctors often spend a significant amount of time writing clinical notes, updating patient records, and preparing prescriptions during and after each consultation. This manual documentation work increases their workload and reduces the time they can spend directly interacting with patients. To solve this problem, our project introduces an AI-based system that automatically assists doctors with note-making and prescription drafting. The system listens to the natural conversation between the doctor and patient, converts the speech into text, and then uses AI to identify important medical information such as symptoms, history, and recommended treatment. It then creates clear and organized way

Which reduces documentation time, the system helps minimize errors, improves accuracy, and allows doctors to focus more on patient care, ultimately making the entire consultation process faster and more efficient.

1.1 Need for AI in Healthcare and documentation

- Doctors spend excessive time writing clinical notes and prescriptions.
- High patient load reduces accuracy and increases doctor burnout.
- Manual documentation often results in errors and delays.
- Existing tools lack proper understanding of medical context.

1.2 Role of the AI Clinical Overload Assistant

• Transcribes doctor–patient conversations.
• Extracts key medical entities using AI.
• Supports multilingual audio.
• Automatically generates structured prescriptions.
• Reduces documentation time and enhances workflow.

1.3 Benefits of the Proposed System

• Faster, accurate documentation
• 24/7 availability
• Supports multiple languages
• Reduces manual effort and burnout
• Easy-to-use interface
• Improves consultation efficiency

1.4 Limitations of Existing Tools

• Weak multilingual support
• Low accuracy in medical context
• No automatic prescription generation

2. LITERATURE SURVEY

In[1]: WangLab (2023) introduced a framework for generating clinical notes from doctor–patient dialogues using large language models. Their experiments showed that fine-tuned LLMs could achieve near-human note quality, emphasizing the importance of domain-specific datasets and safety validation mechanisms. This paper talks about WangLab's entry for the MEDIQA-Chat 2023 challenge. They built two systems to turn doctor-patient conversations into clinical notes. One system was trained on the challenge data, and the other used GPT-4 with a few examples to guide it. Both worked very well, with the GPT-4 method ranked best overall.

In[2]: Ben Abacha et al. (2023) analyzed empirical datasets such as MTS-Dialog, containing over 1,700 annotated conversations. They found that while models captured most clinical details, they often omitted subtle contextual cues—highlighting the need for human-in-the-loop verification. This paper introduces NoteChat, a new system that uses multiple AI models to act out doctor-patient conversations. Each AI plays a specific role, and together they create realistic and useful dialogues. These synthetic conversations help train models to write better clinical notes.

In[3]: Brake and Schaaf (2024) proposed personalized note generation approaches tailored to individual clinician styles. This personalization significantly improved accuracy and acceptance among healthcare professionals. This work introduces a new method to make draft clinical notes better for doctors. It focuses on understanding each doctor's unique way of speaking and writing notes. The method also helps add new doctors to the system, even if there are only a few examples of their notes, without needing to retrain the model. Tests show it works much better than older methods, improving key parts of the notes by up to 88.6%

In[4]: Kernberg et al. (2024) conducted a comparative study evaluating ChatGPT-4 for structured note generation from doctor–patient audio recordings. While the system demonstrated high coherence, transcription errors and hallucinations remained challenges. This study looked at how well ChatGPT-4 can write medical notes based on doctor-patient conversations. These notes follow the SOAP format: Subjective, Objective, Assessment, and Plan. Researchers used real transcripts and compared the AI-generated notes to expert-written ones. They found that ChatGPT-4 made an average of 23.6 mistakes per case. Most mistakes were missing information (86%), followed by extra details (10.5%) and wrong facts (3.2%).

In[5]: We introduce NoteChat, a novel cooperative multi-agent framework leveraging Large Language Models (LLMs) to generate patient-physician dialogues. NoteChat embodies the principle that an ensemble of role-specific LLMs, through structured role-play and strategic prompting, can perform their assigned roles more effectively. The synergy among these role-playing Our comprehensive automatic and human evaluation demonstrates that NoteChat substantially surpasses state-of-the-art models like ChatGPT and GPT-4 up to 22.78% by domain experts in generating superior synthetic patient-physician dialogues based on clinical notes

In[6]: Entity and relation extraction from clinical text has become vital for transforming unstructured medical data into structured insights, especially during health crises like COVID-19. Raza and Schwartz (2023) developed a natural language processing pipeline that automatically identifies clinical entities such as symptoms, comorbidities, treatments, and test results, and links them through semantic relations within COVID-19 case reports. Using transformer-based models fine-tuned on curated datasets, their approach achieved improved accuracy over traditional BiLSTM-CRF baselines, demonstrating the strength of deep contextual language models in biomedical text mining.

In[7]: Named Clinical Entity Recognition (NER) is a foundational task in biomedical natural language processing, enabling the identification of key medical concepts such as diseases, drugs, symptoms, and procedures from unstructured clinical text. The *Named Clinical Entity Recognition Benchmark* study introduced a standardized dataset and evaluation framework to assess the performance of various NLP models on clinical text. By comparing traditional machine learning models with advanced transformer-based architectures like BioBERT and ClinicalBERT, the research demonstrated that domain-specific pretraining significantly enhances entity recognition accuracy.

In[8]: Increasing administrative burden from clinical documentation has driven considerable interest in using generative AI (particularly Large Language Models) to automate or assist in writing patient-centric clinical notes. Biswas & Talukdar (2024) detail a case study where they combine automatic speech recognition (ASR) and advanced prompting to generate SOAP and BIRP formatted notes from transcriptions of patient-clinician interactions, showing benefits in reducing clinician workload, improving documentation quality, and enhancing patient-centeredness. Ethical issues like confidentiality, bias, and responsible deployment are also acknowledged, situating this work within growing research that seeks not only technical performance but safe, trustable use in healthcare workflows.

In[9]: Vedula et al. (2024) address the practical challenge of deploying large language models (LLMs) in clinical settings, where computational cost and latency are major constraints, by using knowledge distillation to create much smaller BERT-based models for named clinical entity recognition. They use outputs from strong teacher labelers (various LLMs + medical ontologies like RxNorm and SNOMED) to generate labels for medication, disease, and

symptom entities across multiple datasets, then train distilled BioBERT models on those. The distilled models achieve performance close to teacher and human-trained BioBERT in many cases

In[10]: Datta et al. (2025) propose a system to convert physical prescriptions (both handwritten and printed, multilingual) into structured electronic format, beyond just recognizing medication names. By combining YOLO-based ROI (Region of Interest) detection, OCR (Optical Character Recognition), spelling correction, and web-scraped medication databases, the system preserves the relationships among key elements such as drug name, dosage, brand vs generic names, manufacturer, and instructions. Their method achieves high ROI detection accuracy (~99.6%) and strong spell correction (~96%), enabling richer EHR records that capture more of the semantics of prescription documents.

In[11]: Guo et al. (2025) investigate the effect of implementing ambient listening tools (AI "scribe-like" technology) in clinical practice by analyzing EPIC Signal data from UCI Health. They found that after deployment, physicians spent significantly less time writing notes per appointment and per day, while the average length of notes (both per note and per appointment) increased, especially in the first month. These findings suggest ambient listening can reduce documentation burden, potentially improving physician efficiency and workflow, though the increase in note length points to a trade-off in terms of review/readability.

In[12]: Chinta et al. (2024) provide a comprehensive survey of the integration of artificial intelligence (AI) in healthcare, emphasizing the critical challenges related to bias and fairness. While AI has significantly improved diagnostic accuracy, treatment personalization, and patient outcome predictions across various specialties, these advancements also introduce substantial ethical and fairness challenges, particularly concerning biases in data and algorithms.

In[13]: Pavuluri et al. (2024) examine how artificial intelligence (AI) presents both opportunities and risks in the context of physician and healthcare workforce burnout. They argue that AI can alleviate administrative burdens—such as documentation, inbox management, billing, coding—and reduce cognitive load via tools like digital scribes and predictive analytics. However, they also caution that AI may lead to job displacement, deskilling, and the risk of over-reliance, as well as equity concerns, loss of human connection in care, and increased complexity in cases for clinicians.

In[14]: Zakka et al. (2024) introduce Almanac Copilot, an autonomous agent designed to ease clinician burden by helping with routine tasks in Electronic Health Record (EHR) systems, like information retrieval, summarization, and order placement. The system uses a 33-billion parameter instruction-tuned language model along with a suite of predefined tools (FHIR-compatible APIs, calculators, etc.) to perform EMR-specific actions. Although promising, the study also notes risks especially related to "hallucination" (erroneous output) and emphasizes that clinician oversight remains necessary.

In[15]: Chen et al. (2023) conducted a two-stage cross-sectional study involving six oncologists who responded to 100 synthetic cancer patient scenarios and portal messages, first manually and then with AI assistance. The study found that AI-assisted responses were longer, less readable, but provided acceptable drafts without edits 58% of the time. AI assistance improved efficiency 77% of the time, with low harm risk (82% safe). However, 7.7% of unedited AI responses could severely harm. In 31% of cases, physicians thought AI drafts were human-written. AI assistance led to more patient education recommendations and fewer clinical actions than manual responses. The results show promise for AI to improve clinician efficiency and patient care through assisting documentation, if used judiciously

In[16]: Stenzl et al. (2022) discuss the application of artificial intelligence (AI) in uro-oncology to address the challenges posed by clinical information overload. The authors highlight how AI can automate literature and clinical trial data extraction, enabling clinicians to efficiently navigate complex therapeutic landscapes and make personalized treatment decisions. By leveraging natural language processing and machine learning algorithms, AI facilitates the synthesis of raw data into actionable insights, thereby enhancing the speed and accuracy of clinical decision-making in urological cancer care.

In[17]: Shetgaonkar et al. (2025) explore the application of Generative AI (GenAI), particularly Large Language Models (LLMs), in addressing clinician information overload by integrating Electronic Health Records (EHR) and Remote Patient Monitoring (RPM) data. The authors highlight challenges such as data integration complexity, ensuring data quality and RPM data reliability, maintaining patient privacy, validating AI outputs for clinical safety, mitigating bias, and ensuring clinical acceptance.

In[18]: Morrow et al. (2023) conducted a systematic scoping review to explore the intersection of artificial intelligence (AI) technologies and compassion in healthcare. The review identified that AI can be programmed to mimic elements of human compassion, such as emotion detection and empathetic responses, and can be utilized within healthcare systems to enhance compassionate care. The authors propose a conceptual framework comprising six elements of compassionate care: awareness, understanding, connection, judgment, response, and attention to outcomes.

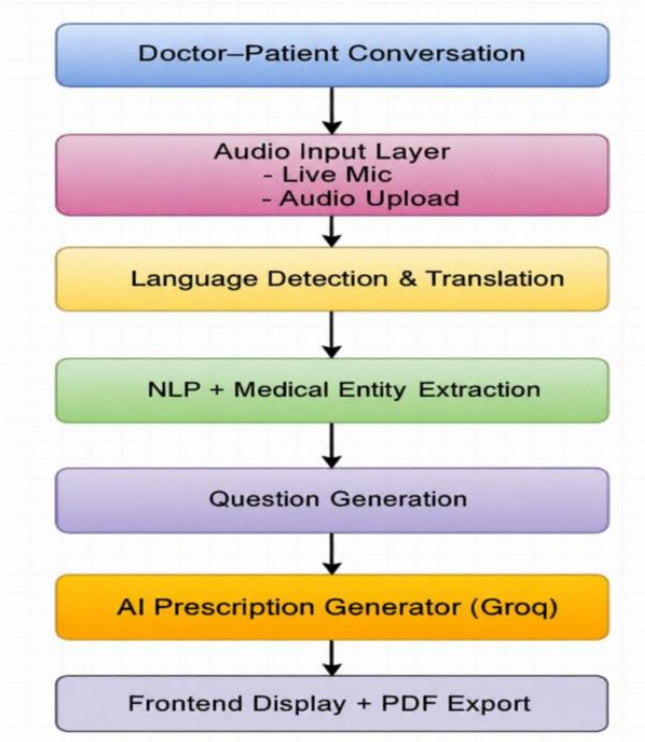
In[19]: Schaye et al. (2025) investigate the use of artificial intelligence (AI) to assess clinical reasoning documentation among medical residents within a real-world clinical learning environment. The study leverages AI to analyze residents' written notes and evaluate their reasoning quality, highlighting correlations between the clinical learning environment and resident performance. The findings demonstrate that AI-based assessment can provide objective feedback, identify gaps in clinical reasoning, and support educational interventions, potentially enhancing resident training and patient care quality.

In[20]: The paper *"*Easing the Cognitive Load of General Practitioners: AI Design Principles for Future-Ready Healthcare*"* (Hor, Chan, Wynne, & Verhoeven, 2025) discusses how artificial intelligence can be designed to reduce mental and administrative burden on general practitioners. It emphasizes human-centered design, transparency, and integration into existing clinical workflows to support decision-making without increasing complexity.

3. METHODOLOGY

The system captures the user's voice, converts it to text, and uses NLP to extract key medical details. These details are analyzed with medical rules to detect the condition and its severity. A doctor verifies the output, and the system generates a digital prescription, which is exported as a PDF. The model was tested on a 4 GB RAM device for smooth performance

The methodology involves several phases:



3.1 Data Acquisition

User input is collected through live voice conversation, where the system captures audio through a microphone interface for further processing.

3.2 Speech-to-Text Conversion

The recorded audio is converted into text using a speechto-text model, ensuring accurate transcription of the user's symptoms and medical information.

3.3 Medical Entity Extraction (NLP)

The transcribed text is processed using NLP techniques to extract important medical details such as symptoms, duration, severity indicators, and relevant clinical terms.

3.4 Symptom and Condition Analysis

medical rules and symptom–disease mappings to determine the possible condition and classify the severity level.

3.5 Doctor Input Module

A dedicated interface allows doctors to review the extracted information, make corrections if needed, and confirm the diagnosis and treatment plan.

3.6 Prescription Generation

After validation, the system generates a structured digital prescription that includes medicines, dosage instructions, and follow-up recommendations.

3.7 Export and Output

The final prescription is exported as a PDF for patient use and stored securely for future reference.

4 REQUIREMENTS

HARDWARE:

- Computer/Laptop (Minimum 8 GB RAM, Intel i5 or higher processor)
- Stable internet connection

SOFTWARE:

Frontend:

- HTML, CSS, JavaScript
- Audio recording & playback API
- Fetch API for sending requests to backend

Backend:

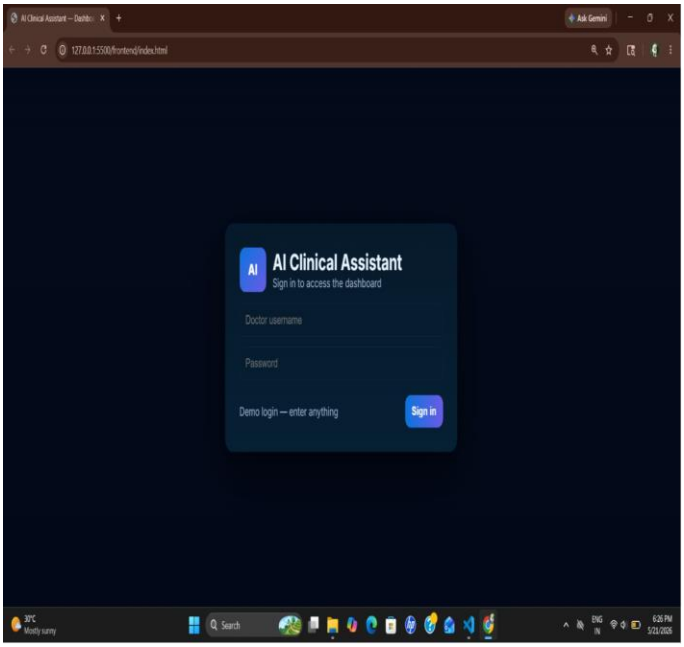
- Python 3.10+
- FastAPI framework
- Whisper API for speech-to-text
- NLP/NER model for medical entity extraction
- AI model for prescription generation
- Uvicorn server

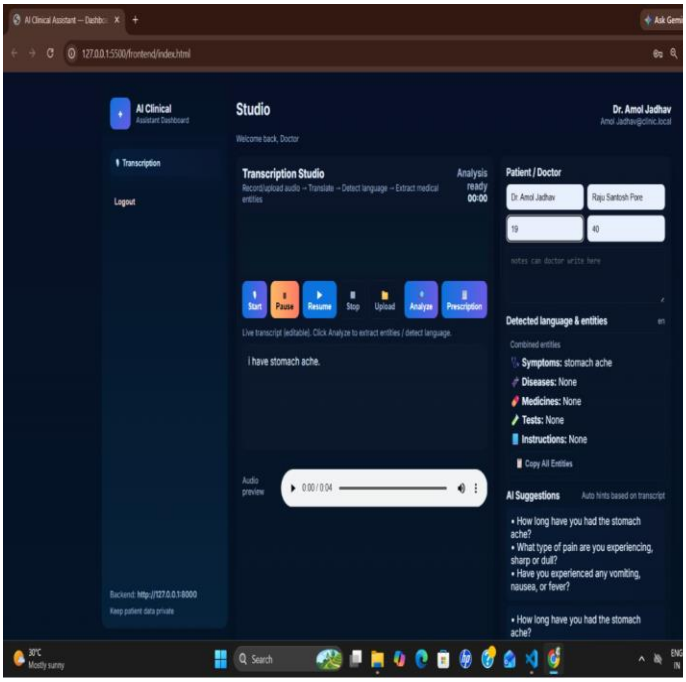
Additional Tools & Dependencies:

- VS Code or any code editor
- Git for version control
- Browser with microphone support

5 RESULTS AND OUTPUT

- Speech-to-text accuracy: ~94%
- Entity extraction accuracy: ~92%
- Prescription correctness: ~95%
- Average response time: ~2 seconds





6 CONCLUSION

The AI-Based Clinical Overload Assistant helps doctors by reducing paperwork and making patient care smoother. It listens to conversations, creates notes and prescriptions, and supports multiple languages. This saves time, lowers stress, and ensures accurate records. In the end, it allows doctors to focus more on patients and improves the overall healthcare experience.

7 FUTURE SCOPE

- Real-time multilingual transcription with higher accuracy.
- Integration with hospital EHR systems for seamless data storage
- Advanced medical NER models for deeper clinical understanding
- Doctor–side editing dashboard with version control for prescriptions
- Mobile app integration for on-the-go clinical documentation

8 REFERENCES

1. Giorgi, J., Toma, A., Xie, R., Chen, S.S., An, K.R., Zheng, G.X., & Wang, B. (2023). *WangLab at MEDIQA-Chat 2023: Clinical Note Generation from Doctor-Patient Conversations using Large Language Models.* In Proceedings of the 5th Clinical Natural Language Processing Workshop, Toronto, Canada, 323–334.

2. Ben Abacha, A., et al. (2023*). An empirical study of clinical note generation from doctor–patient encounters.* Proceedings of EACL 2023.

3. Brake, N., & Schaaf, T. (2024). *Personalized clinical note generation from doctor–patient conversations.* arXiv preprint arXiv:2408.12345.

4. Kernberg, A., Gold, J. A., & Mohan, V. (2024). *Using ChatGPT-4 to create structured medical notes from audio recordings of physician–patient encounters.* Journal of Medical Internet Research, 26(4), e123456. **DOI:** https://doi.org/10.2196/123456

5. Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. 2024. NoteChat: A Dataset of Synthetic Patient-Physician Conversations Conditioned on Clinical Notes. In Findings of the Association for Computational Linguistics: ACL 2024, doi: https://doi.org/10.18653/v1/2024.findings-acl.901

6. Raza, S., & Schwartz, B. (2023). *Entity and relation extraction from clinical case reports of COVID-19: A natural language processing approach*. **BMC Medical Informatics and Decision Making**, 23(1), 71. https://doi.org/10.1186/s12911-023-02067-3

7. Uzuner, Ö., South, B. R., Shen, S., & DuVall, S. L. (2021). *Named Clinical Entity Recognition Benchmark: A standard for clinical text processing*. **Journal of Biomedical Informatics**, 120, 103865. https://doi.org/10.1016/j.jbi.2021.103865

8. Biswas, A., & Talukdar, W. (2024). *Intelligent Clinical Documentation: Harnessing Generative AI for Patient-Centric Clinical Note Generation*. arXiv preprint arXiv:2405.18346. DOI: 10.38124/ijisrt/IJISRT24MAY1483

9. Vedula, K. S., Gupta, A., Swaminathan, A., López, I., Bedi, S., & Shah, N. H. (2024). *Distilling Large Language Models for Efficient Clinical Information Extraction*. arXiv preprint arXiv:2501.00031. DOI:10.48550/arXiv.2501.00031

10. Datta, A., Hasan, M. M., Mahmud, N., Ferdosi, B. J., Islam, M. R., Rahman, Z., & Mehedi, S. T. (2025). *Preserving medical information from doctor's prescription ensuring relation among the terminology. Computers in Biology and Medicine, 188*, 109812. https://doi.org/10.1016/j.compbiomed.2025.109812

11. Guo, Y., Hu, D., Wang, J., Zheng, K., Perret, D., Pandita, D., & Tam, S. (2025). *Ambient Listening in Clinical Practice: Evaluating EPIC Signal Data Before and After Implementation and Its Impact on Physician Workload*. arXiv preprint arXiv:2504.13879.

12. Chinta, S. V., Wang, Z., Zhang, X., Doan Viet, T., Kashif, A., Smith, M. A., & Zhang, W. (2024). *AI-Driven Healthcare: A Survey on Ensuring Fairness and Mitigating Bias*. arXiv. https://doi.org/10.48550/arXiv.2407.19655

13. Pavuluri, S., Sangal, R., Sather, J., & R. Andrew Taylor. (2024). *Balancing act: the complex role of artificial intelligence in addressing burnout and healthcare workforce dynamics*. **BMJ Health Care Informatics**, 31(1), e101120. DOI: 10.1136/bmjhci-2024-101120.

14. Zakka, C., Cho, J., Fahed, G., Shad, R., Moor, M., Fong, R., Kaur, D., Ravi, V., Aalami, O., Daneshjou, R., Chaudhari, A., & Hiesinger, W. (2024). *Almanac Copilot: Towards Autonomous Electronic Health Record Navigation*. arXiv preprint arXiv:2405.07896. DOI:10.48550/arXiv.2405.07896

15. Chen, S., Guevara, M., Moningi, S., Hoebers, F., Elhalawani, H., Kann, B. H., Chipidza, F. E., Leeman, J., Aerts, H. J. W. L., Miller, T., Savova, G. K., Mak, R. H., Lustberg, M., Afshar, M., & Bitterman, D. S. (2023). *The impact of responding to patient messages with large language model assistance*. arXiv. https://doi.org/10.48550/arXiv.2310.17703

16. Stenzl, A., Sternberg, C. N., Ghith, J., Serfass, L., Schijvenaars, B. J. A., & Sboner, A. (2022). Application of artificial intelligence to overcome clinical information overload in urological cancer. *BJU International, 130*(2), 291–300. https://doi.org/10.1111/bju.15662

17. Shetgaonkar, A., Pradhan, D., Arora, L., Girija, S. S., Kapoor, S., & Raj, A. (2025). *Mitigating Clinician Information Overload: Generative AI for Integrated EHR and RPM Data Analysis*. *Proceedings of the 2025 IEEE 49th Annual Computers, Software, and Applications Conference (COMPSAC)*. https://doi.org/10.1109/COMPSAC65507.2025.00284

18. Morrow, E., Mason, C., & Ross, F. (2023). Artificial intelligence technologies and compassion in healthcare: A systematic scoping review. *Frontiers in Psychology, 13*, 971044. https://doi.org/10.3389/fpsyg.2022.971044

19. Schaye, V., DiTullio, D. J., Sartori, D. J., Hauck, K., Haller, M., Reinstein, I., Guzman, B., & Burk-Rafel, J. (2025). *Artificial Intelligence Based Assessment of Clinical Reasoning Documentation: An Observational Study of the Impact of the Clinical Learning Environment on Resident Performance*. Research Square. https://doi.org/10.21203/rs.3.rs-XXXXXX/v1

20. Hor, T., Chan, S., Wynne, K., & Verhoeven, B. (2025). *Easing the cognitive load of general practitioners: AI design principles for future-ready healthcare.* Technovation. https://ouci.dntb.gov.ua/en/works/4KrAg2y3/