# Evaluation of a sequential prompt to measure the time and interactions of a chatbot that schedules medical appointments

Sofia Victoria Pineda Orellana[1], Ashley Priscilla Pérez Yanes[1], Stephanie Abigail Segovia Amaya[1], Wendy Martínez[1], Douglas Aguilar[1], and Jose Luis Ordoñez-Avila[1,2*]
[1]Universidad Evangélica de El Salvador (UEES), El Salvador, douglas.aguilar@uess.edu.sv, wendy.martinez6martinez@uees.edu.sv, jose.ordonez@uees.edu.sv
[2]Universidad Tecnológica Centroamericana (UNITEC), *Honduras,* jlordonez@unitec.edu

***Abstract*– Medical chatbots have raised certain concerns among doctors regarding the level of confidentiality and accuracy of the data. Currently, clinical laboratories have begun integrating a series of non-conversational chatbots to schedule appointments. This paper presents a comparison of conversational chatbots designed to schedule appointments for clinical laboratories. The architecture proposed for the tests included a front end developed in TypeScript, a back end in Node.js, the GPT-4.5 API for conversation, and a Microsoft SQL Server database. Conventional and sequential prompts were evaluated under the same architecture using 45 synthetic scenarios to assess the impact of the proposed sequential prompt. The sequential prompts significantly reduced interaction time and the number of exchanges in both groups of scenarios ($p < 0.05$). Significant effects were observed on interaction time and the number of exchanges ($r = 0.51–0.73$), with the greatest effect observed for the latter. These results indicate that structuring the conversational workflow using sequential prompts can improve the efficiency of chatbots based on large language models (LLMs) for scheduling medical appointments.***

***Keywords*-- Medical chatbot, large language model, prompt engineering, sequential prompt, appointment scheduling.***

## I. Introduction

The healthcare sector has undergone major changes due to the disruptive technologies that have been implemented. Robotics [1], the Internet of Medical Things [2], and artificial intelligence [3] are among the technologies that are evolving in medical centers.

Within these medical centers, clinical laboratories play a fundamental role in disease prevention. These laboratories are difficult to manage due to long waiting times caused by a lack of automation in healthcare systems. In response to this problem, some researchers have made efforts to implement various innovative technologies to optimize these processes. The use of virtual assistants and chatbots on WhatsApp has increased in an attempt to improve service, though without success. These chatbots need new architecture that prevents hallucinations and redundancies and can operate in more natural and reliable environments [4-5],

Therefore, the objective of this study is to design a medical chatbot focused on managing appointments at clinical laboratories using a sequential prompt to streamline interactions.

This research is innovative, first of all, because it incorporates a sequential prompt to keep the ChatGPT API on task. Second, because it compares the sequential prompt with the traditional method of chatbot implementation. And finally, because its architecture ensures security through database encryption, seamless communication via the LLM, and efficiency in performing database queries.

## II. Literary review

### A. Health care security

Medical chatbots based on language models can improve communication and clinical support through personalized and scalable services [5]. Their secure implementation requires governance frameworks and architectures that incorporate privacy by design, data de-identification, federated learning, and secure augmented generation [6]. Likewise, open-source models running locally have demonstrated the potential to provide reliable and empathetic interactions in telemedicine without compromising sensitive information [7]. For these chatbots, patient privacy and reliability are essential characteristics.

The evaluation of medical chatbots must consider the quality of information, the recommendations provided, and user interaction, as improving one component may affect the performance of others [8]. These systems can support education in radiation therapy and continuing education through accessible and personalized experiences [9]. Likewise, the local use of open-source models promotes more private and reliable telemedicine applications, such as the monitoring of patients with hypertension [10]. Their implementation must ensure ethical accountability so that patients feel comfortable engaging in dialogue.

Medical chatbots can expand access to reliable health information and support patient education [11]. This architecture has improved the accuracy of multilingual medication recommendations and the contextualization of responses [12]. Furthermore, combining it with fine-tuning processes enhances performance in clinical consultations, even with limited computational resources [13]. The chatbot only improves consultation performance when paired with a reliable architecture.

To evaluate these systems, a method based on the ISO/IEC 25010 standard was proposed that integrates information quality, clinical guidelines, and clinical process management [8]. It reflects their rapid expansion and raises concerns about

regulation, professional accreditation, transparency, and safety [14]. Although conversational artificial intelligence can enhance communication, diagnostic support, and healthcare, its adoption requires ethical safeguards, privacy protection, and regulatory compliance [14]. Therefore, the growth of these tools must be accompanied by rigorous evaluations to ensure their reliability and safety.

### *B. Large Langue Models*

Advanced language models can improve the efficiency of healthcare services and facilitate more natural, personalized, and context-aware interactions with patients. Their applications include symptom guidance, health education, and clinical decision support [5],[10]. Medical chatbots with RAG architecture combine the natural language understanding and generation capabilities of large language models (LLMs) with information retrieval from structured sources. This integration enables the generation of contextualized, evidence-based responses, improving medication guidance and patient education in areas such as orthopedics [11-13]. However, the processing and retrieval of sensitive clinical data can pose privacy risks, so their implementation requires robust data protection mechanisms [15]. Consequently, the potential of these systems depends on striking a balance between accuracy, clarity, access to reliable knowledge, and data security.

LLMs applied to medical chatbots can understand clinical inquiries and classify user intent when combined with transformer-based models and data balancing techniques [16]. However, processing sensitive information requires privacy-by-design architectures that incorporate de-identification, federated learning, and secure augmented generation [17-18]. Therefore, the development of these systems must integrate accuracy in language understanding, security, and ethical accountability.

Generative AI can transform healthcare through large language models (LLMs) capable of understanding queries, generating context-appropriate responses, and tailoring communication to the patient's needs. These capabilities make it well-suited for use in clinical documentation and diagnosis [19]. However, its implementation requires readable and understandable information, as well as systematic evaluations to ensure the quality of responses [20]. Furthermore, trust, privacy, security, and technological integration are key factors in their acceptance [21].

Privacy and transparency are essential in healthcare chatbots, as they protect sensitive data and promote the ethical use of artificial intelligence [22]. In addition, LLMs enable the generation of context-aware responses and the adaptation of communication style and language to user needs—features that can improve usability, trust, and the effectiveness of the information provided [23-24]. Therefore, their development must combine personalized communication, information protection, and ongoing evaluations.

A review of the literature shows that medical chatbots must incorporate data privacy principles and ensure smooth communication while avoiding hallucinations. To achieve this, it is necessary to propose a modern architecture that connects large language models (LLMs) with sequential processes that ensure interaction with patients.

## III. METHOD

The study evaluated the effect of incorporating a sequential prompt into the design of a chatbot for scheduling medical appointments. Two configurations were compared: a conventional chat and one structured using sequences (Fig. 1). Both used the same architecture, language model, and database, so the prompt design strategy was based on experimentation. The evaluation included 45 synthetic cases organized into two levels of complexity. The first 20 cases provided all the necessary information from the outset to add the data to the database, while the remaining 25 cases included incomplete data, modifications, and additional inquiries, resulting in a greater number of exchanges with the chatbot.

The methodological procedure was carried out in five stages:

1. Definition 1. Definition of scenarios. Twenty synthetic scenarios were prepared in which the user directly enters the information needed to schedule an appointment into the chatbot as part of a conversation. Additionally, 25 scenarios were prepared that require the user to interact with the chatbot to schedule the appointment.
2. Chatbot implementation. The chatbot was configured using the GPT API, and then it was tested with the proposed scenarios. To facilitate comparison, a sequential prompt was added to ensure that all information sent by the user was sufficient to schedule the appointment. In scenarios where information is missing, the chatbot asks the user for the missing data and requests confirmation before scheduling the appointment.
3. Metrics collection. The established variables were recorded to compare the performance of the two configurations. The time taken for the interaction—from when the user initiates the first conversation until they confirm the appointment—was monitored. The number of interactions between the chatbot and the user is also counted, each time the user had to type to schedule the appointment.
4. Statistical analysis. Finally, descriptive statistics, normality tests, and nonparametric tests were performed to determine the impact of using the sequential prompt in the conversation, using time, steps, and their ratio as metrics.

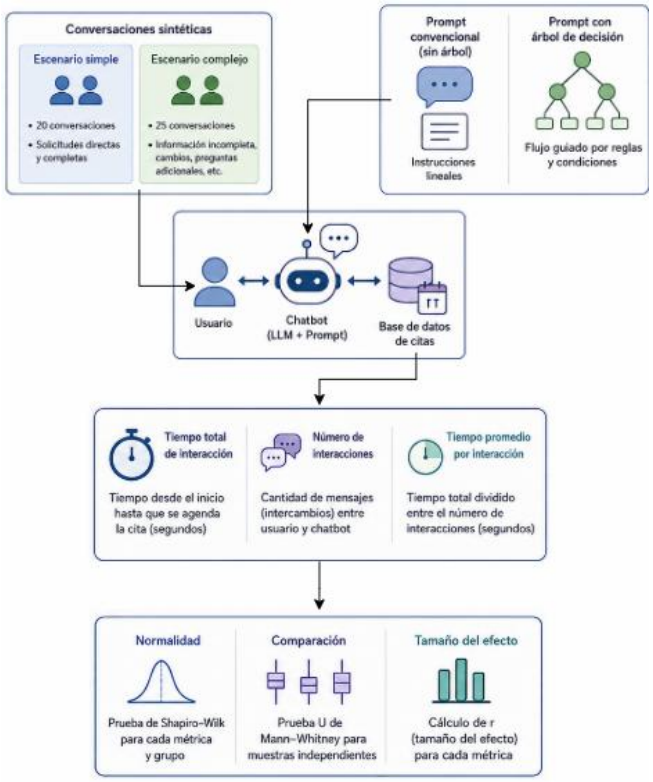


Figure 1. Process for comparison of chatbots strategies.

## A. *Architecture*

The client-server architecture integrated a web interface developed in TypeScript, an artificial intelligence service, and a relational database. Natural language processing was performed using the GPT-4.5 API, with each request built using the configured prompt, available clinical data, and the conversation history. The database was built on Microsoft SQL Server, where patient and appointment information was stored. Additionally, fields were added to track metrics for each session: identifier, date and time, duration, number of interactions, and execution status (Fig. 2).

The backend was organized into layers of routes, controllers, services, and repositories. Routes receive requests from the frontend, while controllers validate and forward the information to the services layer. The *chatbotService* manages the conversational flow, decision tree logic, and metric logging, while the *aiService* builds the context and establishes HTTPS communication with the OpenAI API. Finally, the repositories abstract the read and write operations performed on SQL Server. Both configurations used the same architecture, database, and language model; only the prompting strategy was modified.

## B. *Sequence prompt*

The chatbot was designed using a decision tree-based prompt engineering strategy, with the goal of structuring the medical appointment scheduling process. This strategy organizes the conversation through a sequence of predefined decisions, allowing the chatbot to request only the necessary information at each stage of the process.

The flow begins by identifying the test requested by the user. Once the service is selected, the chatbot proceeds to select the date and time, verifying that the availability matches the schedules recorded in the system.

Next, the chatbot determines who the appointment is for, identifying whether it is for the patient themselves or a minor. Depending on this condition, it requests information from the patient or the legal guardian, ensuring that all data required for registration is collected correctly.

As part of the process, the system offers the option to send confirmation by email. Finally, the chatbot presents a summary of the information entered and requests the user's confirmation before completing the booking process and storing the appointment in the database.

## C. *Statistical Tests*

Once the metrics for each scenario had been collected, a statistical analysis was performed using SPSS. First, the data distribution was assessed using the Shapiro–Wilk test, since the sample size for each group was less than fifty observations. This test made it possible to determine whether the variables analyzed met the assumption of normality required for the application of parametric tests.

Subsequently, since non-normal distributions were identified in most of the variables evaluated, the nonparametric Mann–Whitney U test was used to compare the performance of the chatbot with a conventional prompt and the decision-tree-based chatbot. This test made it possible to determine whether there were statistically significant differences between the two groups for each of the variables analyzed. Finally, in addition to the p-value, the effect size (r) was calculated to quantify the magnitude of the differences found.

Table I. Effect parameters

| Value of (r) | Interpretation | Value of (r) | Interpretation |
|---|---|---|---|
| ($r < 0.10$) | Negligible effect | ($r < 0.10$) | Negligible effect |
| ($0.10 < r < 0.30$) | Small effect | ($0.10 < r < 0.30$) | Small effect |
| ($0.30 < r < 0.50$) | Moderate effect | ($0.30 < r < 0.50$) | Moderate effect |
| ($r > 0.50$) | Large effect | ($r > 0.50$) | Large effect |

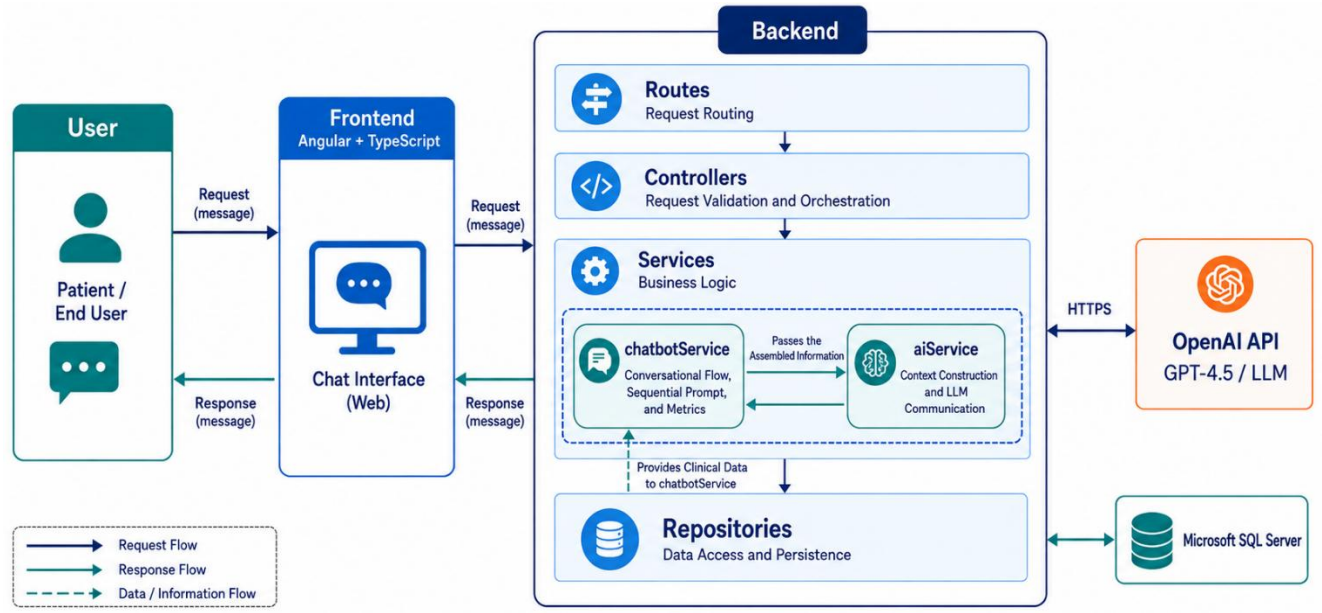


Figure 2. Framework interaction including frontend and baclend.

## IV. Results

Table II shows the descriptive statistics for the synthetic datasets that were used to test the chatbot's performance. These tests were divided into two parts (N=20), where the opening sentence contained all the information needed to schedule an appointment. In N=25, the opening sentence was incomplete, forcing the user to engage in a conversation with the agent. Although the chatbot with sequential prompts has better metrics, it cannot be concluded that it is superior without applying inferential statistics.

Table II. Descriptive statistics data

| **Descriptive** | | Normal | Árbol | Normal | Árbol |
|---|---|---|---|---|---|
| | | N=20 | N=20 | N=25 | N=25 |
| tiempo | Media | 121.51 | 33.31 | 214.38 | 81.39 |
| | Media recortada al 5% | 85.84 | 30.94 | 207.87 | 63.37 |
| | Mediana | 52.43 | 15.31 | 181.94 | 47.42 |
| | Varianza | 34475.41 | 823.63 | 14814.48 | 12207.991 |
| | Desv. Desviación | 185.67 | 28.69 | 121.71 | 110.48 |
| | Mínimo | 22.76 | 8.26 | 51.25 | 13.13 |
| | Máximo | 862.38 | 100.92 | 498.98 | 544.15 |
| Pasos | Media | 4.55 | 3.25 | 7.80 | 4.56 |
| | Media recortada al 5% | 4.50 | 3.22 | 7.82 | 4.56 |
| | Mediana | 4.00 | 3.00 | 8.00 | 5.00 |
| | Varianza | 1.20 | 0.19 | 2.41 | 2.59 |
| | Desv. Desviación | 1.09 | 0.44 | 1.55 | 1.61 |
| | Mínimo | 3.00 | 3.00 | 5.00 | 2.00 |
| | Máximo | 7.00 | 4.00 | 10.00 | 7.00 |
| Pasos / tiempo | Media | 24.34 | 9.76 | 27.09 | 16.23 |
| | Media recortada al 5% | 19.98 | 8.82 | 26.35 | 13.92 |
| | Mediana | 13.37 | 5.10 | 23.2200 | 9.6400 |
| | Varianza | 755.28 | 64.05 | 188.105 | 284.897 |
| | Desv. Desviación | 27.48 | 8.00 | 13.71513 | 16.87890 |
| | Mínimo | 3.88 | 2.75 | 6.41 | 2.82 |
| | Máximo | 123.20 | 33.64 | 62.37 | 77.74 |

The Shapiro–Wilk test revealed non-normal distributions for nearly all variables and configurations evaluated ($p < 0.05$). The only exception was the number of steps in the conventional prompt in the group of 20 cases ($W = 0.906$; $p = 0.053$). Due to the prevalence of non-normal distributions, comparisons were performed using nonparametric tests.

In Scenario 1, the Mann-Whitney U nonparametric test yielded p-values $< 0.05$, indicating that there are significant differences between the two data sets (Table III). When combined with the descriptive statistics, this leads to the conclusion that when the chatbot receives the initial sentence containing all the data, the sequential prompt is more efficient for scheduling the appointment.

Table III. Mann-Whitney U nonparametric test

| | time | steps | steps/time |
|---|---|---|---|

| | | | |
|---|---|---|---|
| U de Mann-Whitney | 81.000 | 57.500 | 109.000 |
| W de Wilcoxon | 291.000 | 267.500 | 319.000 |
| Z | -3.219 | -4.126 | -2.462 |
| Sig. asymptotic(bilateral) | .001 | .000 | .014 |
| Signification [2*(sig. unilateral)] | .001 | .000 | .013 |

Next, the r value is calculated to measure the tree's impact on each of the datasets. The effect size analysis shows that the decision tree had a large impact on the variables time (r = 0.51) and steps (r = 0.65), with the latter showing the greatest magnitude of change. For the steps/time variable, the impact was moderate (r = 0.39). Taken together, these results indicate that the incorporation of the sequential prompt generated significant differences between the groups, with a particularly significant effect on the number of steps required during the interaction.

In Scenario 2, the Mann-Whitney U nonparametric test yielded p-values < 0.05, indicating that the two data sets differ significantly (Table IV). As in Scenario 1, can be concluded that when the chatbot receives the initial sentence with incomplete data, the sequential prompt generates fewer interactions to secure the appointment.

Table IV. Scenario 2Mann-Whitney U nonparametric test

| | time_2 | steps_2 | steps/time |
|---|---|---|---|
| U de Mann-Whitney | 83.000 | 54.500 | 143.000 |
| W de Wilcoxon | 408.000 | 379.500 | 468.000 |
| Z | -4.560 | -5.151 | -3.429 |
| Sig. asymptotic(bilateral) | .000 | .000 | .001 |

The inclusion of the prompt reduced interaction time and the number of exchanges, with large effects (r = 0.64 and r = 0.73, respectively). The relationship between interactions and time showed a moderate effect (r = 0.49). In this case, the prompt had a strong impact on both time and the number of steps, and a technically equivalent impact on the ratio of these two varia

## IV. Conclusions

A framework was designed in Node.js, integrated with a large-scale language model (LLM), to structure questions and control the appointment scheduling workflow. In the overall metrics, the sequential prompt proved to be more efficient than the traditional one. Significant differences were found in both test sets. The strongest effect is observed in the number of interactions, followed by interaction time, suggesting that the decision tree contributes to a more efficient interaction between the user and the chatbot by reducing both the time required and the number of exchanges needed to complete the interaction.

## References


[1] J. L. Ordoñez-Avila y M. Cardona, «Chapter 9 - Internet of things challenges for medical solutions», en Rehabilitation Robotics and Healthcare Devices, M. Cardona, Ed., Academic Press, 2025, pp. 185-194. doi: https://doi.org/10.1016/B978-0-443-21505-6.00002-5.

[2] M. Cardona, J. L. Ordoñez-Avila, y F. Serrano, «Chapter 1 - A categorization of medical robots by their applications», en Rehabilitation Robotics and Healthcare Devices, M. Cardona, Ed., Academic Press, 2025, pp. 1-12. doi: 10.1016/B978-0-443-21505-6.00008-6.

[3] K. A. Castro, J. A. Siwady, E. Castillo, A. Alonzo, M. Cardona and M. E. Perdomo, "Artificial Intelligence for All: Challenges and Harnessing Opportunities in AI Democratization," 2024 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT), San Salvador, El Salvador, 2024, pp. 143-148, doi: 10.1109/ICMLANT63295.2024.00030.

[4] J. L. Ordoñez-Avila, I. Mendez, M. E. Perdomo, J. F. I. Dubón, K. S. M. López and J. H. R. Agüero, "Microclasses Gamification Test for Students Using an Android Robot as an Assistant," 2026 IEEE Engineering Education World Conference (EDUNINE), Mexico City, Mexico, 2026, pp. 1-6, doi: 10.1109/EDUNINE62390.2026.11547042.

[5] J. L. Ordoñez-Avila, M. E. Perdomo, J. F. I. Dubón, K. S. M. López, J. H. R. Agüero and I. Mendez, "Level of Confidence in Using an Android Robot as an Assistant in Robotics Microclasses," 2026 IEEE Engineering Education World Conference (EDUNINE), Mexico City, Mexico, 2026, pp. 1-6, doi: 10.1109/EDUNINE62390.2026.11546906.

[6] Chow, J. C. L., Wong, V., & Li, K. (2024c). Generative Pre-Trained Transformer-Empowered Healthcare Conversations: Current Trends, Challenges, and Future Directions in Large Language Model-Enabled Medical Chatbots. BioMedInformatics, 4(1), 837-852.

[7] Kim, E. J., & Kim, J. (2025b). Privacy-by-Design Framework for Large Language Model Chatbots in Urology. International Neurourology Journal, 29(Suppl 2), S65-S72. https://doi.org/10.5213/inj.2550274.137

[8] Montagna, S., Ferretti, S., Klopfenstein, L. C., Ungolo, M., Pengo, M. F., Aguzzi, G., & Magnini, M. (2025b). Privacy-preserving LLM-based chatbots for hypertensive patient self-management. Smart Health, 36, 100552. https://doi.org/10.1016/j.smhl.2025.100552

[9] Barletta, V. S., Caivano, D., Colizzi, L., Dimauro, G., & Piattini, M. (2023). Clinical-chatbot AHP evaluation based on "quality in use" of ISO/IEC 25010. International Journal of Medical Informatics, 170, 104951. https://doi.org/10.1016/j.ijmedinf.2022.104951

[10] Magnini, M., Aguzzi, G., & Montagna, S. (2025). Open-source small language models for personal medical assistant chatbots. Intelligence-Based Medicine, 11, 100197. https://doi.org/10.1016/j.ibmed.2024.100197

[11] Chow, J. C. L., & Li, K. (2025i). Large Language Models in Medical Chatbots: Opportunities, Challenges, and the Need to Address AI Risks. Information, 16(7), 549. https://doi.org/10.3390/info16070549

[12] Baur, D., Ansorg, J., Heyde, C.-E., & Voelker, A. (2025c). Development and Evaluation of a Retrieval-Augmented Generation Chatbot for Orthopedic and Trauma Surgery Patient Education: Mixed-Methods Study. JMIR AI, 4(1), e75262. https://doi.org/10.2196/75262

[13] Bora, A., & Cuayáhuitl, H. (2024b). Systematic Analysis of Retrieval-Augmented Generation-Based LLMs for Medical Chatbot Applications. Machine Learning and Knowledge Extraction, 6(4), 2355-2374. https://doi.org/10.3390/make6040116

[14] El Bazzi, M. S., Anass, T., & Abdelouahed, A. (2025). Advancing Multilingual Retrieval-Augmented Generation for Reliable Medication Counseling. IEEE Access, PP, 1-1. https://doi.org/10.1109/ACCESS.2025.3646941

[15]Chu, B., Modi, N. D., Menz, B. D., Bacchi, S., Kichenadasse, G., Paterson, C., Kovoor, J. G., Ramsey, I., Logan, J. M., Wiese, M. D., McKinnon, R. A., Rowland, A., Sorich, M. J., & Hopkins, A. M. (2025a). Generative AI's healthcare professional role creep: A cross-sectional evaluation of publicly accessible, customised health-related GPTs. Frontiers in Public Health, 13. https://doi.org/10.3389/fpubh.2025.1584348

[16]Guan, S., Kwok, H. C., Law, N. F., Stiglic, G., Qin, H., & Hui, V. (2025). Privacy Challenges and Solutions in Retrieval-Augmented Generation-Enhanced LLMs for Healthcare Chatbots: A Review of Applications, Risks, and Future Directions (arXiv:2511.11347). arXiv. https://doi.org/10.48550/arXiv.2511.11347

[17]Assayed, S. K., Shieh, C.-S., & Gupta, S. K. (2025d). Engineering intelligent healthcare systems: Understanding medical queries with AI and NLP. Journal of Engineering and Applied Science, 72(1), 212. https://doi.org/10.1186/s44147-025-00800-y

[18]Kim, E. J., & Kim, J. (2025b). Privacy-by-Design Framework for Large Language Model Chatbots in Urology. International Neurourology Journal, 29(Suppl 2), S65-S72. https://doi.org/10.5213/inj.2550274.137

[19]Feretzakis, G., Anastasiou, A., Pitoglou, S., Paxinou, E., Gkoulalas-Divanis, A., Kalodanis, K., Tsapelas, I., Kalles, D., & Verykios, V. S. (2024). Securing a Generative AI-Powered Healthcare Chatbot. Studies in Health Technology and Informatics, 321, 195-199. https://doi.org/10.3233/SHTI241091

[20]Rabbani, S. A., El-Tanani, M., Sharma, S., Rabbani, S. S., El-Tanani, Y., Kumar, R., & Saini, M. (2025). Generative Artificial Intelligence in Healthcare: Applications, Implementation Challenges, and Future Directions. BioMedInformatics, 5(3), 37. https://doi.org/10.3390/biomedinformatics5030037

[21]Olszewski, R., Brzeziński, J., Watros, K., & Rysz, J. (2026a). Quantifying Readability in Chatbot-Generated Medical Texts Using Classical Linguistic Indices: A Review. Applied Sciences, 16(3). https://doi.org/10.3390/app16031423

[22]Wah, J. N. K. (2025b). Revolutionizing e-health: The transformative role of AI-powered hybrid chatbots in healthcare solutions. Frontiers in Public Health, 13. https://doi.org/10.3389/fpubh.2025.1530799

[23]Yener, R., Chen, G.-H., Gumusel, E., & Bashir, M. (2025a). Can I Trust This Chatbot? Assessing User Privacy in AI-Healthcare Chatbot Applications. https://doi.org/10.48550/arXiv.2509.14581

[24]Koscelny, S. N., Sadralashrafi, S., & Neyens, D. M. (2025). Generative AI responses are a dime a dozen; Making them count is the challenge – Evaluating information presentation styles in healthcare chatbots using hierarchical Bayesian regression models. Applied Ergonomics, 128, 104515. https://doi.org/10.1016/j.apergo.2025.104515