Conversational Medical AI: Ready for Practice

Anonymous submission

Abstract

The shortage of doctors is creating a critical squeeze in access to medical expertise. While conversational Artificial Intelligence (AI) holds promise in addressing this problem, its safe deployment in patient-facing roles remains largely unexplored in real-world medical settings. We present the first large-scale evaluation of a physician-supervised LLM-based conversational agent in a real-world medical setting.

Our *agent* was integrated into an existing medical advice chat service. Over a three-week period, we conducted a randomized controlled experiment with 926 cases to evaluate patient experience and satisfaction. Among these, *The Agent* handled 298 complete patient interactions, for which we report physician-assessed measures of safety and medical accuracy.

Patients reported higher clarity of information (3.73 vs 3.62 out of 4, $p < 0.05$) and overall satisfaction (4.58 vs 4.42 out of 5, $p < 0.05$) with AI-assisted conversations compared to standard care, while showing equivalent levels of trust and perceived empathy. The high opt-in rate (81% among respondents) exceeded previous benchmarks for AI acceptance in healthcare. Physician oversight ensured safety, with 95% of conversations rated as "good" or "excellent" by general practitioners experienced in operating a medical advice chat service.

Our findings demonstrate that carefully implemented AI medical assistants can enhance patient experience while maintaining safety standards through physician supervision. This work provides empirical evidence for the feasibility of AI deployment in healthcare communication and insights into the requirements for successful integration into existing healthcare services.

1 Introduction

Globally, persistent shortages and inequitable distribution of the health workforce contribute to decreased access to health services and poorer quality of care. Projections indicate a shortage of 10 million health workers worldwide by 2030 (Boniol et al. 2022). Countries across Europe are facing shortages in primary care physicians, aggravated by aging populations and increased chronic disease burden (Russo et al. 2023). Regional disparities are particularly pronounced, with urban areas generally having higher physician densities than rural regions (Winkelmann, Muench, and Maier 2020; Pál et al. 2021). Studies report deteriorating access to care, especially in these underserved areas, leading to increased workloads and burnout among practitioners (Dumesnil et al. 2024). Physician burnout is associated with reduced engagement and lower quality of care (Zhou et al. 2020). The limited availability of primary care services not only restricts access to preventive and routine care, but also creates additional strain on emergency services, ultimately degrading the overall quality of care (Russo et al. 2023).

Recent advances in general-purpose large language models (LLMs) and generative AI have opened new opportunities for healthcare applications, particularly through conversational AI agents optimized for medical use (Tu et al. 2024). Such agents can serve a number of critical roles fundamental to a patient's care, health literacy, coordination, and management. By directly answering patients' medical questions more readily, collecting relevant diagnostic information, and facilitating patient-provider communication, they could help address the growing challenges in access and quality of care. This potential has prompted active research into the safety, accuracy, and effectiveness of conversational AI agents in healthcare settings.

Retrospective and modeling analyses show that AI agents perform increasingly well on metrics evaluating diagnostic accuracy, answers to patient-directed medical questions, knowledge recall, and medical reasoning (Tu et al. 2024; Zeltzer et al. 2023; Singhal et al. 2023b,a). In Tu et al. (2024), AMIE (Articulate Medical Intelligence Explorer), an LLM-based AI system optimized for clinical history-taking and diagnostic dialogue, demonstrated greater diagnostic accuracy and superior performance compared to physicians in simulated consultations with patient actors (Tu et al. 2024). Evaluating the safety and performance of patient-facing conversational AI agents in a real-world setting is among the next steps forward.

A health and insurance company operating in multiple European countries (referred to as *The Company*) , has offered a medical chat advice service to its members since 2020. Using *The Company's* mobile app, any member can ask a question directly to an on-call physician through the privacy-compliant chat. In 2024, *The Company* introduced *The Agent*, an LLM-based conversational agent, to this medical advice chat service staffed by its general practitioners.

In this study, we present our findings from this experiment in introducing conversational AI into medical practice.

Our primary contributions are:

- We introduced *The Agent* , a patient-facing medical agent designed as an AI system. To this end, we developed a comprehensive evaluation framework combining clinical knowledge and reasoning assessment, real-world conversation analysis, and automated testing through simulated patient interactions.
- We integrated *The Agent* into a pre-existing medical advice chat service, with a focus on ethical design for patients, physician oversight, and quality assurance.
- We ran a randomized controlled experiment, collecting data over 3 weeks to compare patient satisfaction and experience between conversations when *The Agent* was proposed and a control group of patients that interacted solely with human physicians. The experiment highlighted that overall satisfaction and perceived clarity were higher in conversations with *The Agent* , while trust in the received information and perceptions of empathy were similar between the two groups. We also show that patient engagement is higher in conversations with *The Agent* , evidenced by shorter response times from patients.
- We evaluated safety and medical accuracy through physician reviews. 95% of the conversations were assessed as "good" or "excellent", while no conversation was considered as potentially dangerous overall.
- Finally, we discussed the implications of our findings for the broader adoption of AI in healthcare, focusing on patient empowerment, access to care, and the evolution of healthcare delivery models.

2 *The Agent* , an LLM-based medical conversational agent deployed in *The Company*'s medical chat

2.1 Context

The Company is a health services and insurance firm, providing health coverage for seven hundred thousand members as of October 2024.

In 2020, *The Company* introduced a medical advice chat service as a way to enhance its product and service offerings for its members. Using the mobile app, members can directly contact a general practitioner or specialist physician to receive answers to their medical questions during extended hours (from 7 am to 12 am, seven days a week). The medical advice chat service is fully compliant with health privacy regulations, and uses end-to-end encryption for the messages between members and physicians.

Between January 1 and October 1, 2024, *The Company*'s medical advice chat service facilitated over 58,000 conversations between members and health professionals. These conversations were split between general practitioners (62%) and other healthcare professionals specializing in physiotherapy, nutrition, gynecology, pediatrics, dermatology and sexual health. At the beginning of the study, general practitioners (GPs) had been operating the service for an average of 2.8 years (range: 0.8 - 4.0). Towards supporting the doctors operating the service, *The Company* introduced an LLM-based conversational agent into its medical advice chat service over the summer of 2024.

2.2 Developing *The Agent*, an LLM-based Medical Conversational Agent

Objective The objective of *The Company*'s conversational AI agent is to provide users with clear, appropriate, and actionable responses to their medical and healthcare questions while maintaining positive rapport and trust.

A Multi-Agent Systemic Approach Rather than a single, standalone LLM, *The Agent* is an LLM-based AI system consisting of several sub-agents that run in parallel. This multi-agent systemic approach allows *The Agent* to use the best model for each specific task, integrating the strengths of different models within the system (Rasal and Hauer 2024; Guo et al. 2024; Shen et al. 2024). The models are served in compliance with EU privacy regulations.

Design Process and Offline Evaluation To design *The Agent*'s system architecture and select its constituent LLMs, we developed a comprehensive offline evaluation framework comprising three components: (i) a clinical knowledge and reasoning benchmark of 800+ multiple-answer closed questions from a national medical exam, (ii) anonymized past conversations from the medical advice chat for testing response quality, and (iii) simulated conversations with patient agents to evaluate complete end-to-end interactions. This framework allowed us to assess *The Agent*'s performance across medical knowledge, reasoning, and communication capabilities while enabling evaluation of rare or difficult scenarios through simulation.

2.3 Integrating *The Agent* into the medical advice chat service

A product team of engineers, designers, doctors, and user researchers collaborated to integrate *The Agent* into the medical advice chat service in a safe, intuitive, and transparent way. *The Agent* was deployed between 9 am and 11 pm for conversations addressed to GPs, with patients who consented to automated treatment of their data.

Ethical Compliance
We established comprehensive guidelines to ensure ethical compliance. We anticipated the entry into force of the EU AI Act (European Parliament and Council of the European Union 2024), augmenting its recommendations to ensure responsible implementation and a transparent interface that patients can easily understand.

To ensure responsible AI deployment, we implemented the following safeguards: (1) timely human review consisting in physician oversight (2) explicit and implicit (e.g., color of text bubbles) differentiation between AI agents and human actors, (3) consent collection for health data processing using LLMs, (4) requiring positive action for interaction with *The Agent* (see Supplementary Figure S2) , and (5) clearly limiting the scope of conversations for which *The Agent* can operate. For example, in cases of psychological emergency, *The Agent* was inactivated.

Physician Oversight
The Agent operates under the supervision and responsibility of the physicians of the medical advice chat service.

Physician-agent interface. GPs have the authority and capability to stop *The Agent* and intervene during any patient-agent conversation, regardless of whether *The Agent* is composing a message or waiting for the patient to reply. *The Agent* never resumes the conversation once stopped. The GP is required to check in with the patient after the exchange between *The Agent* and the patient is complete.

Message review. As a conversation between a patient and *The Agent* unfolds, a GP assigned to the conversation is required to review each message from *The Agent* within 15 minutes. GPs can hide *The Agent*'s messages when necessary. Hiding a message requires the GP to take over the discussion, and displays the message in a "hidden" state to the patient while keeping it visible to the GP. In cases of urgency, GPs can immediately establish direct contact with patients using their provided contact information.

General conversation review. If *The Agent* has been involved in a conversation, the assigned GP must perform a general review. This review consists of examining the complete *The Agent*-patient dialogue to evaluate the medical advice provided and identify any potential gaps or concerns. The GP then documents their assessment and engages directly with the patient for a mandatory check-in to confirm their oversight, validate *The Agent* 's medical recommendations, provide complementary guidance when needed, and address any remaining questions (see Supplementary Figure S2) .

Staged Roll-out and Quality Assurance
The Agent's deployment progressed through three sequential stages over a four-month period ending in October 2024. The first stage limited access to *The Company*'s employees only, allowing for initial validation. The service was then extended to a small proportion of members under the supervision of GPs selected and trained to support *The Agent*'s development. Finally, access was expanded to 50% of members with oversight from all GPs of the medical advice chat service after they received specific training. Each stage lasted as long as necessary to reach defined safety and stability milestones.

Throughout the integration, a team of physicians and engineers continuously monitored safety and stability metrics established during development, enabling data-driven improvements while maintaining rigorous quality standards.

3 Methods

3.1 Study Design

We conducted a randomized controlled experiment to evaluate the effect of our LLM-based conversational agent, on patient experience. From all eligible conversations, *The Agent* was proposed to a random 50% sample of patients to comprise the treatment group, while the remaining eligible conversations served as the control group. We evaluated patient experience across three domains: (i) overall satisfaction, (ii) quality metrics (clarity, trust, and empathy), and (iii) engagement metrics (response patterns).

In addition to assessing patient experience, we evaluated *The Agent*'s safety and medical accuracy from the physician message and general conversation reviews.

Data was prospectively collected from September 30 to October 20, 2024.

3.2 Outcome measures

We developed questionnaires to evaluate both the patient experience of conversations with *The Agent* and the physician assessments of safety and accuracy of *The Agent*'s responses. To do so, we surveyed existing standards for evaluation of patient-doctor interactions (PACES exam (of the Royal Colleges of Physicians of the UK 2023), GMC Patient Questionnaire (Council 2024), Best Practice for Patient Centered Care (King and Hoppe 2013)) and extracted core information on our specific domains of interest. We differentiated between patient-related outcomes to be reported by the patient and medical assessment to be conducted by a physician, while considering constraints in length and user experience to maximize completion rate.

Patient Ratings
Following each conversation, patients were asked to rate their experience across four dimensions: overall satisfaction, clarity, trust, and empathy (see Supplementary Table S1, Supplementary Figure S5). Information on patient satisfaction was captured using a 5-point Likert scale and free text. Clarity, trust, and empathy were assessed using a 4-point Likert scale.

GP General Review
After each complete *The Agent*-patient conversation, the assigned GP evaluated its quality. They assessed *The Agent*'s questioning, recommendations, and accuracy, and also provided an overall assessment of the conversation. All used a 4-point Likert scale apart from accuracy, which was rated on a 3-level scale (see Supplementary Table S2, Supplementary Figure S6).

Statistical Analysis and Consent for Research
We compared distributions of patient and GP ratings using the Wilcoxon test. Demographic comparisons were conducted using Student's t-test for age and chi-squared test for gender. All statistical analyses were conducted using R version 4.3.1.

We excluded from the study all conversations with attachments (document, picture) and conversations with *The Company*'s employees. Data from conversations requesting unavailable services (prescriptions, sick leave certificates, or medical certificates) were excluded from the patient experience analysis.

All members included in this study were informed of the use of aggregated and/or anonymized data for research and statistical purposes in *The Company*'s Privacy Policy. Additionally, members provided explicit consent for the automated processing of their health data using LLM technology.

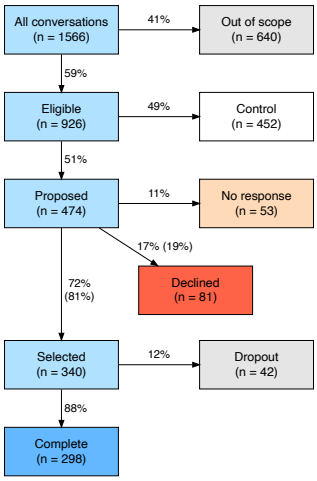

Figure 1: **Flow diagram of *The Agent* deployment in medical advice conversations**. Of 1,566 conversations where *The Agent* was active, 640 (41%) were out of scope. Among eligible conversations (n = 926), *The Agent* was proposed to 474 patients, with 452 as controls. After excluding no-responses (n = 53) and declines (n = 81), 340 patients opted to interact with *The Agent*, of whom 298 (88%) completed their conversations. Percentages in parentheses represent rates adjusted for no-responses.

4 Results

4.1 Sample Profile

Over the study period, 1,566 conversations were initiated in *The Company*'s medical advice chat service during *The Agent*'s active hours (Figure 1). *The Agent* deemed 640 conversations (41%) out of scope, due to questions that contained insurance or administrative matters or signs of mental health distress that, by established protocols, required human intervention.

Of the 926 eligible conversations, *The Agent* was proposed to 474 patients (51%), while 452 conversations served as the control group. Among those offered *The Agent*, 53 patients (11%) did not respond within the required 15-minute window before GP takeover, likely because they expected an asynchronous response and were not actively monitoring their chat. Of the remaining patients who responded, 81 (19%) declined interaction, resulting in 340 patients opting to interact with *The Agent*, an acceptance rate of 81% among respondents.

Among those who began interacting with *The Agent*, 298 patients (88%) completed their conversations, while 42 patients (12%) dropped out before completion as assessed by the monitoring physician.

The demographic characteristics across conversation categories are presented in Supplementary Table S3. The mean age of users across all conversations was 34.5 years, with a higher proportion of female users (63%). Among eligible conversations, the control and *The Agent* Proposed groups showed comparable demographic profiles (mean age difference: 0.4 years [95% CI: -0.5 to 1.4]; difference in female proportion: -0.7% [95% CI: -7.0% to 5.6%]). The demographic characteristics in completed conversations (mean age: 32.6 years, 68% female) remained consistent with the initial eligible population (mean age difference: 0.5 years [95% CI: -0.6 to 1.5]; difference in female proportion: 1.3% [95% CI: -5.0% to 7.6%]).

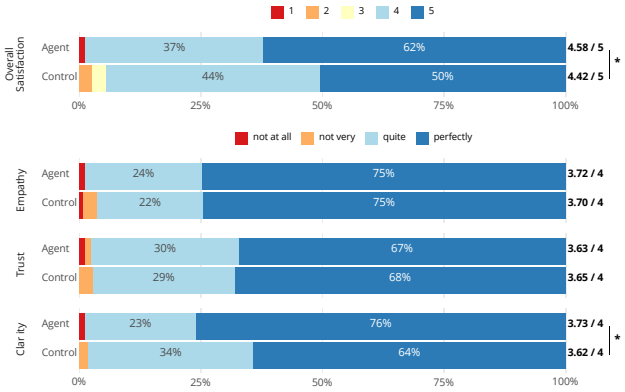

Figure 2: **Patient ratings:** comparison between *Agent* and control groups. **Top:** Overall satisfaction rated on a 5-point scale (1:😠, 5:😻). **Bottom:** Specific dimensions (Empathy, Trust, Clarity) rated on a 4-point scale ('not at all' to 'perfectly'). Numbers on the right show mean scores. Asterisks (*) indicate statistically significant differences between groups ($p < 0.05$).

4.2 Patient Experience

Patient ratings were available for 20% of eligible conversations. Ratings were more prevalent in the control group (24% vs 17%), and demographic characteristics were comparable between the two groups (mean age difference: 1.6 years [95% CI: -0.8 to 3.9]; difference in female proportion: -3% [95% CI: 11% to 17%]).

The Agent received higher general satisfaction scores compared to the control group (mean: 4.58 vs 4.42 out of 5, $p < 0.05$) (Figure 2). Both treatment and control groups showed similar ratings for trust (mean: 3.63 vs 3.65 out of 4) and empathy (mean: 3.72 vs 3.70 out of 4). However, *The Agent* achieved significantly higher clarity ratings (mean: 3.73 vs 3.62 out of 4, $p < 0.05$).

Notably, extremely low ratings (score of 1) were rare. *The Agent* received only one such rating across all dimensions, and the control group received one rating of 1 for empathy only. A detailed analysis of all ratings below 3 (n = 8) revealed no systematic patterns of dissatisfaction (Supplementary Table S4) .

4.3 Safety and Medical Accuracy

GPs supervising the medical advice chat service evaluated *The Agent*'s performance at both message and conversation levels (Figure 3) . At the message level, supervising GPs reviewed each of *The Agent*'s responses within 15 minutes of sending. Among 1,265 messages sent by *The Agent*, 95% were rated positively, while 45 messages (3.6%) were rated

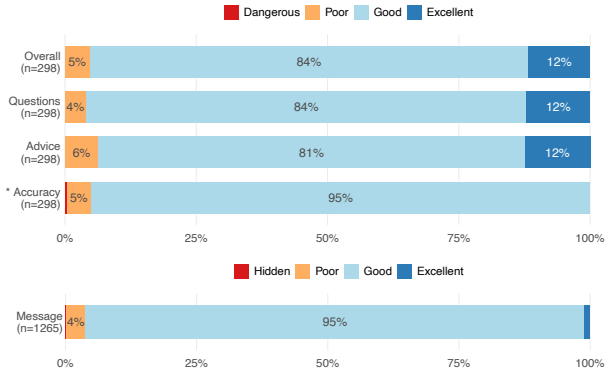

Figure 3: **GP evaluation of *The Agent*'s medical quality at message and conversation levels. Top:** Conversation-level assessment (n=298) across different dimensions. Each conversation was evaluated for overall performance, quality of questions asked, advice given, and accuracy. Ratings range from "dangerous" (red) to "excellent" (dark blue), except for Accuracy (*) which was rated specifically for presence of inaccuracies (none/some/dangerous). **Bottom:** Message-level review (n=1,265) of individual responses from *The Agent*, rated from "hidden" (red) to "excellent" (dark blue).

as "poor" and 3 messages were hidden from patients. No harm resulted from the messages that were subsequently hidden from patient view.

Following the completion of each conversation, GPs provided an overall assessment. For completed conversations (n=298), 95% received positive ratings ("good" or "excellent") for overall performance, with similar distributions for question quality (96%) and advice appropriateness (94%). No conversation was deemed potentially dangerous overall.

In the assessment of medical accuracy, 95% of conversations contained no inaccuracies, with one conversation flagged for the presence of potentially dangerous inaccuracies.

5 Discussion

5.1 Bridging AI Research and Clinical Practice

This section describing prior work has been abridged to fit.

Noteworthy is the novelty of the study, which deploys conversational medical AI at scale, in real-world conditions, for the first time. It mainly builds on previous research that looked at real-world deployment at smaller scale (n = 26, 19, 34, (Dwyer et al. 2023), (Yang et al. 2024), (Jo et al. 2023)), and previous research that evaluated AI systems in simulated environments ((Tu et al. 2024), (Ayers et al. 2023)).

5.2 Understanding Patient Experience: Satisfaction, Trust, and Engagement

Implications for Healthcare Delivery

Building on these promising but limited pilots, our study presents the first large-scale deployment of an AI medical assistant in a real-world healthcare setting, with close to 300 completed patient conversations. Our findings on patient satisfaction merit careful interpretation within the broader context of healthcare delivery. Patient satisfaction is a crucial prerequisite for broader acceptance and adoption of AI in healthcare. The comparable or superior satisfaction ratings achieved in conversations with *The Agent* indicates the feasibility of AI deployment in clinical settings. This acceptance could enable significant reconfiguration of healthcare delivery systems, potentially allowing for more efficient allocation of human medical expertise while maintaining or improving access to care. Specifically, AI agents could evolve into daily health companions, fundamentally shifting healthcare from episodic interventions to continuous support, where patients are empowered to better understand and manage their health journey, while being efficiently connected to physician expertise when needed.

Dimensions of Patient Satisfaction

The granular analysis of satisfaction metrics reveals important nuances in patient experience. The significantly higher clarity ratings suggest that AI-assisted communications may excel at providing clear, structured information, aligning with previous findings that standardized communication approaches can enhance patient understanding (Trevena et al. 2005).

The equivalent ratings for trust and empathy warrant particular attention. Unlike studies where raters were unaware of AI involvement (e.g., (Tu et al. 2024; Ayers et al. 2023)), our transparent setup explicitly identified *The Agent* as an AI agent. Previous research on AI interactions suggests that perceived humanness increases feelings of trust and empathy (Lu et al. 2022; Hu, Lu, and Gong 2021). Therefore, the comparable ratings are especially significant given that knowledge of *The Agent*'s AI status could have influenced patient expectations. Two factors likely contributed to maintaining trust despite transparent AI use: *The Agent*'s consistent responsiveness and structured communication style, and our protocol ensuring that a physician personally engages with the patient at the end of each conversation.

Patient Engagement and Communication Dynamics

The high opt-in rate (81% among respondents) indicates strong patient acceptance of AI-assisted healthcare services, setting a higher benchmark for user acceptance than previously suggested in the literature (Horowitz et al. 2023; Esmaeilzadeh, Mirzaei, and Dharanikota 2021). Through user interviews, we identified three factors potentially contributing to this success: (i) members' trust in *The Company*, built over time (ii) an iteratively refined user experience, and (iii) an emphasis on transparency.

Analysis of conversation dynamics (see Supplementary) revealed that *The Agent*'s nearly instantaneous responses were associated with faster patient response times. These accelerated exchanges could fundamentally improve healthcare delivery. Fluid dialogue leads to more comprehensive information gathering, while rapid response times could lower the barrier to seeking medical advice, encouraging patients to address health concerns earlier. This aligns with previous research showing that reduced response latency can enhance user engagement and satisfaction in healthcare communications (Yang et al. 2024; Wu et al. 2023). The

combination of AI responsiveness and physician oversight creates a new model where patients benefit from both immediate attention and expert medical judgment.

These findings suggest that successful integration of AI in healthcare services depends not only on technical capabilities but also on careful attention to user experience, institutional trust, and transparent implementation practices. The results demonstrate that when properly implemented, AI-assisted healthcare services can achieve high levels of patient acceptance while maintaining high quality standards in medical communication.

5.3 Ethical, Privacy, and Safety concerns of AI-based Communication Systems for Health

From a safety perspective, the results of our study are encouraging yet warrant careful consideration. While 95% of *The Agent*'s messages received positive physician reviews and only three messages (out of 1,265) required intervention, the few cases where mitigation was required by the supervising GP confirms the need for physician oversight in this setup and continued research. In particular, extended data collection will allow observation of a broader range of rare cases that may elicit inappropriate responses from *The Agent*.

Earlier studies emphasized several prerequisites for deploying patient-facing AI systems in healthcare: stringent quality control measures, sufficient guardrails, adequate oversight by qualified physicians, ethical design and development, as well as strict adherence to privacy regulations and informed consent procedures (Wu et al. 2023; Busch et al. 2024; Haltaufderheide and Ranisch 2024; Meskó and Topol 2023). The integration of *The Agent* in *The Company*'s medical advice chat demonstrates a practical realization of these requirements in a real-world healthcare setting.

The following steps were critical in ensuring its reliability. First, we established comprehensive offline evaluation procedures, comprising of: (i) the constitution of an internal closed-questions benchmark, tailored to the needs relevant to the deployment of *The Agent*, and unlikely to be used in the prior training of the LLMs we use, (ii) the use of anonymized past conversation data representative of the specific task, and (iii) the development of an automated conversation evaluation framework involving patient agents. Second, we carefully integrated *The Agent* in the final product, insisting on (i) the thoughtful design of the interaction between the physician and *The Agent*, prioritizing physician oversight and leveraging user experience to elicit the right actions (e.g., timely message review), and (ii) a staged rollout to enable learning and iterations before full-scale implementation.

This study was made possible by two critical aspects of our development process. First, we build upon a pre-existing medical service. Second, *The Agent* and its integration into the patient-facing product were developed by a multidisciplinary team that included a dedicated GP, aligning with recommendations made by others (Zhou et al. 2024).

5.4 Study Limitations

This real-world evaluation has several important limitations. The three-week duration may not capture seasonal variations in health issues or longer-term patterns in patient-AI interactions. More importantly, while substantial for an initial deployment, this sample size may not be sufficient to detect rare but significant safety issues that could emerge in broader medical practice.

The evaluation of patient experience was constrained by our survey response rate of 20%. While this rate is typical for embedded product surveys, it introduces potential selection bias in our satisfaction metrics. Despite finding no significant demographic differences between respondents and non-respondents, there may be unmeasured factors influencing survey participation that correlate with patient satisfaction.

Our study scope was also limited in several practical ways. We restricted *The Agent*'s deployment to general practitioner conversations, excluding consultations with other specialists, which might present different challenges. The exclusion of conversations requiring document review or image analysis, while necessary for our initial deployment, leaves important use cases unexplored. Additionally, as the study was conducted within a single healthcare system with an established digital presence, our findings about patient acceptance may not generalize to other healthcare contexts, particularly those without pre-existing patient trust in digital services.

5.5 Future Research Priorities

Our study demonstrates the potential of AI-assisted medical communication while highlighting key areas for future research. Longer-term studies should examine how AI assistance affects healthcare delivery and outcomes, including impacts on patient health-seeking behavior, quality of preventive care, and physician workload. Research should focus on optimizing the collaboration between AI systems and healthcare professionals, establishing efficient oversight models, and developing protocols for seamless care transitions. Technical advances such as integration with electronic health records and capabilities for handling medical documents and images would enable more comprehensive care support. Additionally, continued research into improving the handling of complex medical presentations and rare conditions remains essential for reliable deployment at scale.

6 Conclusion

Our findings demonstrate the feasibility and far-reaching potential of AI-assisted medical communication, while highlighting the importance of careful implementation and oversight. The success of this implementation relied heavily on the integration of medical expertise throughout development, robust privacy protections, and continuous safety monitoring. While results are promising, longer-term studies with larger sample sizes are needed to fully understand the impact of AI-assisted medical communication on healthcare delivery, access and quality of care, and patient outcomes.

A Supplementary Figures

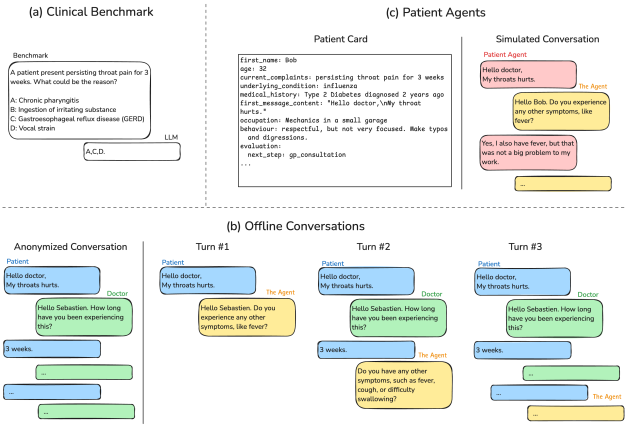

Figure S1: **Offline evaluation methods. (a)** Multiple-choice medical exam questions assess medical knowledge and clinical reasoning. **(b)** Real-world medical advice conversations evaluate response quality and relevance. **(c)** Simulated conversations with patient agents evaluate end-to-end information gathering and recommendation accuracy.

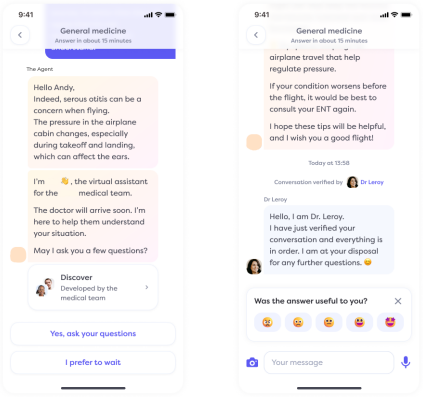

Figure S2: **Transparent user interface. Left:** When patients initiate a conversation in the medical advice chat, *The Agent* first reformulates their concern and explicitly asks for their preference: they can either start with *The Agent*'s assistance or opt to wait for a physician. **Right:** At the end of *The Agent* interactions, physicians engage directly with the patient to acknowledge their oversight of the conversation, validate *The Agent*'s medical guidance, and provide complementary advice when necessary. Here, we also show the entry point for the user ratings survey.

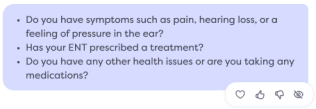

Figure S3: **Physician review interface for *The Agent* messages.** Physicians review each message and select one of the four rating icons within 15 minutes. The right-most choice removes the message from the patient's view.

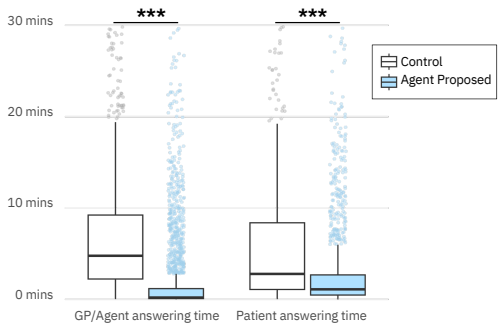

Figure S4: **Response time distributions in medical chat conversations.**
Left: Time taken by providers to respond (*The Agent* or GP) after the patient. **Right:** Time taken by patients to respond. Box plots show median, interquartile range, and whiskers (1.5 IQR); individual points represent outliers beyond whiskers. The visualization is cropped on the Y axis. In the *The Agent* Proposed group, patients interact with both *The Agent* and the GP. Asterisks (***) indicate statistically significant differences ($p < 0.001$).
Commentary: As expected, since *The Agent* responds almost instantaneously, response times from providers differed significantly (median: 0.2 vs 4.8 minutes, $p < 0.001$). Interestingly, this difference in provider response times was accompanied by a change in patient behavior: in conversations with *The Agent*, patients also responded more quickly compared to control conversations (median: 1.1 vs 2.8 minutes, $p < 0.001$).

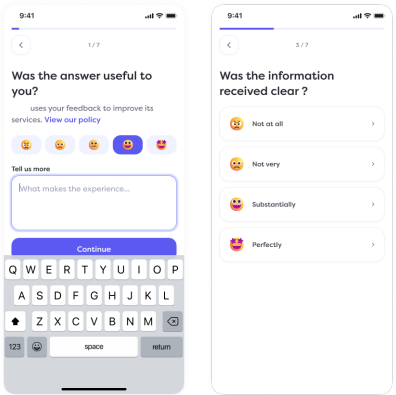

Figure S5: Example screens for member feedback

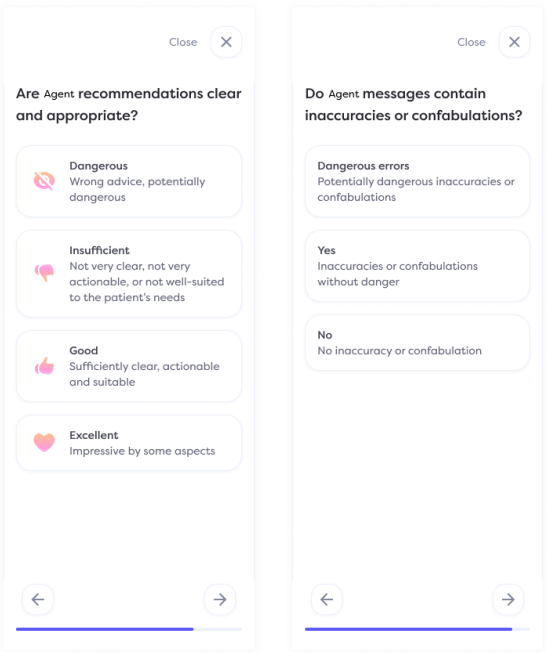

Figure S6: Example screens for physician evaluation

B Supplementary Tables

Table S1: **Patient experience questionnaire**

Category	Question	Rating Scale
Overall satisfaction	How useful was the conversation?	😡 😞 😐 😀 😍
Clarity	How clear was the information you've received?	Not at all; Not very; Substantially; Perfectly
Trust	How much do you trust the information you've received?	Not at all; Not very; Substantially; Perfectly
Empathy	How heard and understood have you felt?	Not at all; Not very; Substantially; Perfectly

Table S2: **General practitioner assessment**

Category	Question	Assessment
Advice	Are *The Agent*'s recommendations clear and appropriate?	**Dangerous:** Wrong advice, potentially dangerous **Insufficient:** Not very clear, not very actionable, or not well-suited to the patient's needs **Good:** Sufficiently clear, actionable and suitable **Excellent:** Impressive by some aspects
Questions	Are *The Agent*'s questions relevant and well-phrased?	**Dangerous miss:** Essential questions are missing or poorly phrased **Insufficient:** Some missing or poorly phrased questions **Good:** Sufficient questions posed **Excellent:** Perfect! No unnecessary questions.
Accuracy	Do *The Agent*'s messages contain inaccuracies or confabulations?	**Dangerous errors:** Potentially dangerous inaccuracies or confabulations **Yes:** Inaccuracies or confabulations without danger **No:** No inaccuracy or confabulation.
Overall Assessment	Overall, the conversation between *The Agent* and the patient seemed to you...	**Dangerous** **Laborious** **Satisfactory** **Amazing**

Table S3: **Demographic characteristics by conversation status and group.** Age and gender distribution across conversation categories. Age is presented as mean and range [25th - 75th percentiles], with minimum and maximum values. F prop. represents the proportion of conversations with female users. Groups are mutually exclusive and follow the flow diagram (Figure **??**).

		Age		Gender		
	Conversations	Mean	min [q25 - q75] max	Female	Male	F prop.
All Conversations	1,566	34.5	17 [28 - 39] 72	983	575	63%
Eligible	**926**	**32.1**	**18 [27 - 36] 67**	**619**	**302**	**67%**
Control	452	31.9	18 [27 - 35] 67	304	146	68%
***The Agent* Proposed**	**474**	**32.3**	**18 [27 - 36] 64**	**315**	**156**	**67%**
The Agent No Answer	53	33.4	18 [29 - 36] 64	34	19	64%
The Agent Declined	81	31.0	18 [26 - 34] 55	48	32	60%
***The Agent* Selected**	**340**	**32.5**	**18 [27 - 36] 63**	**233**	**105**	**69%**
Dropout	42	31.7	20 [26 - 36] 53	29	11	72%
Complete	**298**	**32.6**	**18 [27 - 36] 63**	**204**	**94**	**68%**

Table S4: **Details of poorly rated conversations.** We show here all conversations with a poor rating. Overall Satisfaction: below 3/5; Clarity, Trust and Empathy: below 2/4. Impact of *The Agent* on negative ratings seems limited.

Group	Description	Role of *The Agent*	Overall Satisfaction	Clarity	Trust	Empathy
Control	Low rating justified. Patient asks a clear pediatric question and the physician makes a diagnosis too quickly without answering the initial question.	Not involved	2	2	2	2
Control	Low rating partly justified. Doctor is not assessing the problem because a GP is on their way to do a physical examination, and it's the best solution. GP could have been more empathetic and pedagogic.	Not involved	2	3	3	2
Control	Medium rating without apparent justification. The answer was great, and the patient seemed happy with the conversation.	Not involved	3	3	3	3
Control	Low rating justified. Patient came for psychological distress and was not redirected or provided with options.	Not involved	3	2	2	1
Control	Medium rating without apparent justification. Patient came for a complaint that needed further examination and was invited to consult in real life.	Not involved	3	4	4	4
Control	Low rating justified. Patient is concerned about their daughter. Doctors advised calling emergency services (15) without asking more questions or giving advice.	Not involved	2	3	2	2
The Agent Proposed	Medium rating without apparent justification. Patient asked for pediatric advice; *The Agent* answered well, and the doctor validated the response.	Good behavior	4	3	2	3
The Agent Proposed	Low rating partly justified. Patient asked for an appointment with a specialist for a chronic issue (3 years). They requested the phone number of our doctors but were redirected to teleconsultation.	*The Agent* promised help to find a specialist but failed to give useful advice, which might have annoyed the member.	1	1	1	1

References

Ayers, J. W.; Poliak, A.; Dredze, M.; Leas, E. C.; Zhu, Z.; Kelley, J. B.; Faix, D. J.; Goodman, A. M.; Longhurst, C. A.; Hogarth, M.; and Smith, D. M. 2023. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Internal Medicine*, 183(6): 589.

Boniol, M.; Kunjumen, T.; Nair, T. S.; Siyam, A.; Campbell, J.; and Diallo, K. 2022. The global health workforce stock and distribution in 2020 and 2030: a threat to equity and 'universal' health coverage? *BMJ Global Health*, 7(6): e009316.

Busch, F.; Hoffmann, L.; Rueger, C.; van Dijk, E. H.; Kader, R.; Ortiz-Prado, E.; Makowski, M. R.; Saba, L.; Hadamitzky, M.; Kather, J. N.; Truhn, D.; Cuocolo, R.; Adams, L. C.; and Bressem, K. K. 2024. Systematic Review of Large Language Models for Patient Care: Current Applications and Challenges. *medRxiv*.

Council, G. M. 2024. Patient feedback (or feedback from those you provide medical services to). https://www.gmc-uk.org/registration-and-licensing/managing-your-registration/revalidation/guidance-on-supporting-information-for-revalidation/patient-feedback---or-feedback-from-those-you-provide-medical-services-to.

Dumesnil, H.; Lutaud, R.; Bellon-Curutchet, J.; Deffontaines, A.; and Verger, P. 2024. Dealing with the doctor shortage: a qualitative study exploring French general practitioners' lived experiences, difficulties, and adaptive behaviours. *Family Practice*.

Dwyer, T.; Hoit, G.; Burns, D.; Higgins, J.; Chang, J.; Whelan, D.; Kiroplis, I.; and Chahal, J. 2023. Use of an Artificial Intelligence Conversational Agent (Chatbot) for Hip Arthroscopy Patients Following Surgery. *Arthroscopy, Sports Medicine, and Rehabilitation*, 5(2): e495–e505.

Esmaeilzadeh, P.; Mirzaei, T.; and Dharanikota, S. 2021. Patients' Perceptions Toward Human–Artificial Intelligence Interaction in Health Care: Experimental Study. *Journal of Medical Internet Research*, 23(11): e25856.

European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 March 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. OJ L 90, 15.3.2024. Available at: http://data.europa.eu/eli/reg/2024/1689/oj.

Guo, T.; Chen, X.; Wang, Y.; Chang, R.; Pei, S.; Chawla, N. V.; Wiest, O.; and Zhang, X. 2024. Large Language Model based Multi-Agents: A Survey of Progress and Challenges. arXiv:2402.01680.

Haltaufderheide, J.; and Ranisch, R. 2024. The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). *npj Digital Medicine*, 7(1).

Horowitz, M. C.; Kahn, L.; Macdonald, J.; and Schneider, J. 2023. Adopting AI: how familiarity breeds both trust and contempt. *AI & SOCIETY*, 39(4): 1721–1735.

Hu, P.; Lu, Y.; and Gong, Y. Y. 2021. Dual humanness and trust in conversational AI: A person-centered approach. *Computers in Human Behavior*, 119: 106727.

Jo, E.; Epstein, D. A.; Jung, H.; and Kim, Y.-H. 2023. Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, 1–16. ACM.

King, A.; and Hoppe, R. B. 2013. "Best Practice" for Patient-Centered Communication: A Narrative Review. *Journal of Graduate Medical Education*, 5(3): 385–393.

Lu, L.; McDonald, C.; Kelleher, T.; Lee, S.; Chung, Y. J.; Mueller, S.; Vielledent, M.; and Yue, C. A. 2022. Measuring consumer-perceived humanness of online organizational agents. *Computers in Human Behavior*, 128: 107092.

Meskó, B.; and Topol, E. J. 2023. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *npj Digital Medicine*, 6(1).

of the Royal Colleges of Physicians of the UK, F. 2023. PACES - MRCP(UK) Part 2 Clinical Examination. https://www.thefederation.uk/examinations/paces.

Pál, V.; Lados, G.; Ilcsikné Makra, Z.; Boros, L.; Uzzoli, A.; and Fabula, S. 2021. Concentration and inequality in the geographic distribution of physicians in the European Union, 2006-2018. *Regional Statistics*, 11(3): 3–28.

Rasal, S.; and Hauer, E. J. 2024. Navigating Complexity: Orchestrated Problem Solving with Multi-Agent LLMs. arXiv:2402.16713.

Russo, G.; Perelman, J.; Zapata, T.; and Šantrić Milićević, M. 2023. The layered crisis of the primary care medical workforce in the European region: what evidence do we need to identify causes and solutions? *Human Resources for Health*, 21(1).

Shen, W.; Li, C.; Chen, H.; Yan, M.; Quan, X.; Chen, H.; Zhang, J.; and Huang, F. 2024. Small LLMs Are Weak Tool Learners: A Multi-LLM Agent. arXiv:2401.07324.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; Payne, P.; Seneviratne, M.; Gamble, P.; Kelly, C.; Babiker, A.; Schärli, N.; Chowdhery, A.; Mansfield, P.; Demner-Fushman, D.; Agüera y Arcas, B.; Webster, D.; Corrado, G. S.; Matias, Y.; Chou, K.; Gottweis, J.; Tomasev, N.; Liu, Y.; Rajkomar, A.; Barral, J.; Semturs, C.; Karthikesalingam, A.; and Natarajan, V. 2023a. Large language models encode clinical knowledge. *Nature*, 620(7972): 172–180.

Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Hou, L.; Clark, K.; Pfohl, S.; Cole-Lewis, H.; Neal, D.; Schaekermann, M.; Wang, A.; Amin, M.; Lachgar, S.; Mansfield, P.; Prakash, S.; Green, B.; Dominowska, E.; y Arcas, B. A.; Tomasev, N.; Liu, Y.; Wong, R.; Semturs, C.; Mahdavi, S. S.; Barral, J.; Webster, D.; Corrado, G. S.; Matias, Y.; Azizi, S.; Karthikesalingam, A.; and Natarajan, V. 2023b. Towards Expert-Level Medical Question Answering with Large Language Models. arXiv:2305.09617.

Trevena, L. J.; (Hons), H. M. D. B. H. M.; Barratt, A.; Butow, P.; and Caldwell, P. 2005. A systematic review on communicating with patients about evidence. *Journal of Evaluation in Clinical Practice*, 12(1): 13–23.

Tu, T.; Palepu, A.; Schaekermann, M.; Saab, K.; Freyberg, J.; Tanno, R.; Wang, A.; Li, B.; Amin, M.; Tomasev, N.; Azizi, S.; Singhal, K.; Cheng, Y.; Hou, L.; Webson, A.; Kulkarni, K.; Mahdavi, S. S.; Semturs, C.; Gottweis, J.; Barral, J.; Chou, K.; Corrado, G. S.; Matias, Y.; Karthikesalingam, A.; and Natarajan, V. 2024. Towards Conversational Diagnostic AI. arXiv:2401.05654.

Winkelmann, J.; Muench, U.; and Maier, C. B. 2020. Time trends in the regional distribution of physicians, nurses and midwives in Europe. *BMC Health Services Research*, 20(1).

Wu, C.; Xu, H.; Bai, D.; Chen, X.; Gao, J.; and Jiang, X. 2023. Public perceptions on the application of artificial intelligence in healthcare: a qualitative meta-synthesis. *BMJ Open*, 13(1): e066322.

Yang, Z.; Xu, X.; Yao, B.; Rogers, E.; Zhang, S.; Intille, S.; Shara, N.; Gao, G. G.; and Wang, D. 2024. Talk2Care: An LLM-based Voice Assistant for Communication between Healthcare Providers and Older Adults. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(2): 1–35.

Zeltzer, D.; Herzog, L.; Pickman, Y.; Steuerman, Y.; Ber, R. I.; Kugler, Z.; Shaul, R.; and Ebbert, J. O. 2023. Diagnostic Accuracy of Artificial Intelligence in Virtual Primary Care. *Mayo Clinic Proceedings: Digital Health*, 1(4): 480–489.

Zhou, A. Y.; Panagioti, M.; Galleta-Williams, H.; and Esmail, A. 2020. *Burnout in Primary Care Workforce*, 59–72. Springer International Publishing. ISBN 9783030609986.

Zhou, H.; Liu, F.; Gu, B.; Zou, X.; Huang, J.; Wu, J.; Li, Y.; Chen, S. S.; Zhou, P.; Liu, J.; Hua, Y.; Mao, C.; You, C.; Wu, X.; Zheng, Y.; Clifton, L.; Li, Z.; Luo, J.; and Clifton, D. A. 2024. A Survey of Large Language Models in Medicine: Progress, Application, and Challenge. arXiv:2311.05112.