

Towards Fluid Consultations: Design and Real-World Evaluation of a Multi-Agent, Full-Duplex Health AI Assistant

Zhoujian Sun, Minghui Yang, Jian Wang, Jinjie Gu, Junwei Liu

Ant Group

Correspondence: baishun@antgroup.com

Abstract

Access to reliable health guidance is a critical global challenge. While large language models show promise, their practical utility is typically limited by high input barriers, lack of adaptive response strategies for complex scenarios, and insufficient real-world evaluation. To address these gaps, we present a novel health AI assistant that integrates a low-latency full-duplex voice interface with a Multi-Agent System (MAS). In our within-user observational analysis, full-duplex use was associated with greater conversation depth (10.16 vs. 2.62 turns for text), but the non-randomized modality assignment precludes a causal interpretation. Furthermore, our MAS architecture decouples clinical reasoning from general conversation, effectively managing complex queries and boosting clinical performance scores by approximately 30% compared with the single-agent baseline. We conducted the first large-scale analysis of an LLM-based health assistant deployed to the general public ($N = 5,000$), demonstrating superior user engagement and providing empirical evidence on the utility of deploying such systems in realistic environments.

1 Introduction

Access to reliable health information remains a global challenge. Individuals seeking guidance for health conditions often face a fragmented and untrustworthy information landscape (Chou et al., 2018). These barriers are amplified for sensitive topics: people with mental or sexual health concerns may avoid timely help from communities or medical professionals due to stigma (Arias López et al., 2023; Pretorius et al., 2019). Disparities in health literacy and socioeconomic status further compound the problem, leaving vulnerable populations least equipped to find and evaluate credible guidance (Mackert et al., 2016). Such health inequities impose an estimated annual economic burden of billions of dollars (LaVeist et al., 2023),

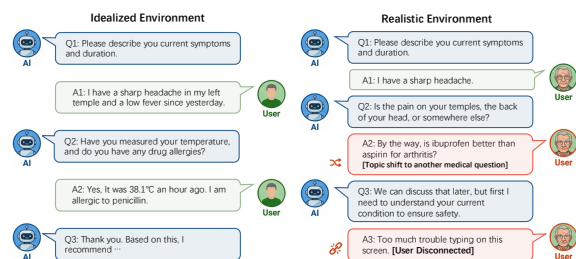


Figure 1: User Behavior Differences Between Different Environments. In idealized settings, users exhibit perfect compliance in answering assistant questions (left). However, they exhibit complex non-compliance behaviors such as topic shifting and premature termination in realistic settings (right).

underscoring the need for trustworthy and accessible health support tools (Laranjo et al., 2018).

The advent of large language models (LLMs) offers an unprecedented opportunity to provide broad access to health artificial intelligence (AI) assistants (Thirunavukarasu et al., 2023). Trained on vast textual corpora, LLMs can support multi-turn dialogue, synthesize medical knowledge, and generate natural responses (Chen et al., 2025). Recent studies show that LLM-based assistants can achieve performance comparable to human doctors across specialized medical tasks (Sellersgren et al., 2025), including passing medical licensing examinations (Singhal et al., 2023), conducting preliminary diagnoses through conversation (Johri et al., 2025), and analyzing complex patient medical records (Small et al., 2025).

However, these results are largely derived from idealized environments. Existing research often overlooks the complexity of real-world deployment scenarios (Figure 1) and fails to address challenges inherent to practical application (Choudhury et al., 2025; Liu et al., 2025). This gap limits the broader population from benefiting from LLM advances. In unconstrained environments, current LLM-based agents face substantial interactional limitations.

First, accessibility remains a critical bottleneck (High Input Barriers). Existing AI health interfaces mostly rely on text input or rudimentary half-duplex voice interaction, creating barriers for populations that may need healthcare support most, including older adults, visually impaired users, and those with limited digital literacy. Typing detailed symptoms into a chat interface can be difficult or impossible, while conventional voice systems often lack interruption handling and natural turn-taking. As noted by (Zhang et al., 2024), cumbersome typing or rigid voice interaction may reduce users’ willingness to provide detailed information, limiting the context needed for accurate responses.

Second, current models often employ rigid conversational strategies. In clinical contexts, the optimal interaction style depends on user intent: simple questions may require concise answers, whereas complex scenarios, such as a sudden, severe headache, require proactive clarification before suggestions are given (Tu et al., 2025). Yet LLMs fine-tuned primarily on general instruction-following data often struggle with this distinction. They tend to adopt a “one-size-fits-all” pattern, either passively waiting for prompts or producing overwhelming blocks of text, rather than switching adaptively between concise consultant and thorough interviewer.

Finally, rigorous evaluation in real-world scenarios remains lacking. Most studies focus on static medical datasets, and although patient simulators are used as proxies, their credibility has not been rigorously validated against real patient interactions characterized by ambiguity and emotional distress. Without data from actual public use, real-world risks related to safety and hallucination remain underexplored (Team et al., 2025; Fan et al., 2025; Schmidgall et al., 2026; Sviridov et al., 2025).

In this study, we present the first comprehensive design and large-scale real-world evaluation of a user-centric health AI assistant. We introduce three targeted improvements to bridge laboratory prototypes and medical utility. **Full-Duplex Voice Interaction:** We developed a real-time voice architecture that supports natural turn-taking; in observational logs, its use was associated with an increase in average interaction depth from 2.62 turns for text to 10.16 turns. **Multi-Agent Orchestration:** To address rigid response strategies, we implemented a Multi-Agent System (MAS) that dynamically routes requests to specialized agents,

such as Clinical Inquiry and Search, improving performance on complex clinical queries by approximately 30% (from 0.68 to 0.88). **Real-World Evaluation:** We analyzed data from 5,000 public users to characterize engagement patterns and identify critical safety benchmarks for future deployment.

2 System Implementation

Our assistant uses a cascaded pipeline to support natural, interruptible, and reliable human-AI interaction (Figure 2).

2.1 Audio Interface Implementation

We used the Silero Voice Activity Detection (VAD) model (Team, 2021) to monitor user speech at 16 kHz with 32 ms chunks, producing a speech probability score. Based on preliminary experiments, we set the detection threshold to 0.6. For *barge-in*, the system immediately stops text-to-speech (TTS) generation when VAD detects user speech during assistant output. Specifically, it enters the listening state when the VAD probability exceeds the threshold for five consecutive chunks, and determines speech has ended only after 800 ms without valid speech, preventing premature cutoffs.

For Automatic Speech Recognition (ASR) and TTS, we used FireRedASR (Xu et al., 2025) and Index-TTS-2 (Zhou et al., 2025), respectively. Both models were optimized for streaming inference. TTS begins synthesis after receiving the first five tokens from the agents, keeping audio processing overhead negligible.

The ASR transcript, together with the dialogue history, is submitted to the MAS to generate a response. Since most existing MAS are not designed for real-time interaction and often require tens of seconds, we set a 0.5-second engineering target for MAS inference latency rather than a latency guarantee. To meet this constraint, we designed the following framework.

We adopted a two-layer MAS. The first layer is a planner that analyzes the dialogue and routes the current request to an appropriate agent without directly responding. The second layer contains independent agents that handle user requests in detail. In each turn, only one agent is activated. We avoided common designs such as multi-agent debate because of the latency constraint. We used MoE LLMs for both the planner and agents, as activating only a small subset of parameters per

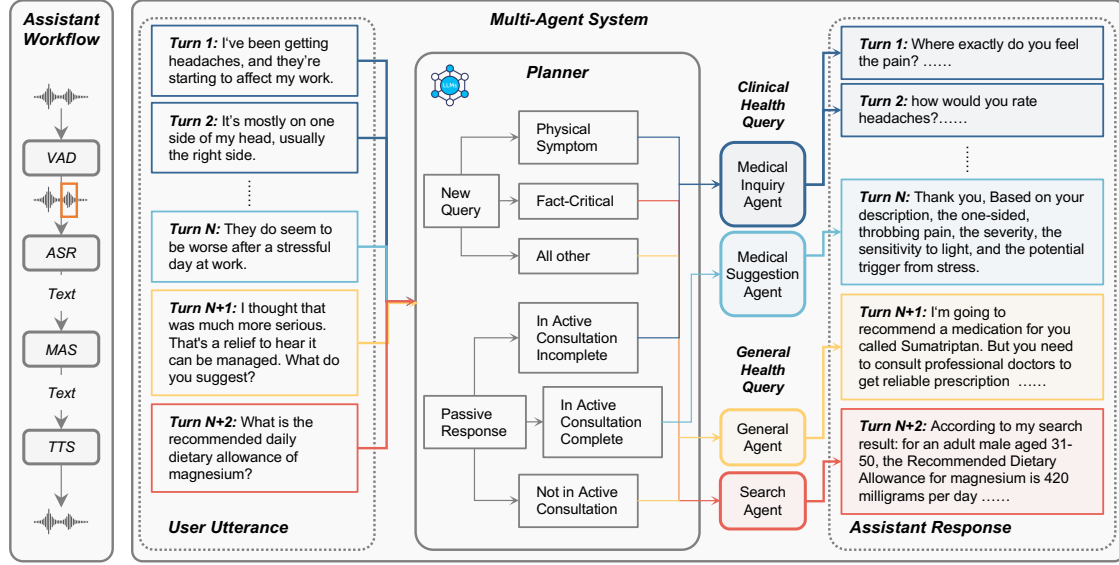


Figure 2: System Implementation. In the assistant workflow, the voice activity detection (VAD) model continuously detects the audio, segmenting meaningful human speech (the sections in the orange boxes) from the continuous audio stream. The segmented audio chunks are then sent to the ASR model, which transcribes the audio into text. In the MAS, the planner processes the textual input and dialogue history to select the most appropriate agent, which subsequently generates the specific output. Finally, the text-to-speech (TTS) model synthesizes the agent’s textual output into audio during each interaction.

token enables faster inference. Given the strict latency requirement, we did not use LLMs with Chain-of-Thought reasoning. The planner uses Qwen3-30B-A3B-Instruct-2507, which activates only 3 billion parameters per token and can select the downstream agent in about 0.2s via in-context learning (Yang et al., 2025). Each agent is based on Qwen3-235B-A22B-Instruct-2507 (Yang et al., 2025) and generates tokens in a streaming manner.

We carefully designed the agent roles. Although health-related conversations can be topically complex, user queries can generally be divided into two types (Heydari et al., 2025). The first type, **clinical health queries**, involves assistant-led dialogue in which the user responds passively (Li et al., 2025; Fansi Tchango et al., 2022; Johri et al., 2025). Typical examples include diagnostic interviews and intervention recommendations, where the user presents an initial concern and the doctor-like assistant then collects information across multiple turns before giving a final suggestion (Sun et al., 2024). The second type, **general health queries**, is user-led, with the assistant responding passively, as in general QA scenarios (Brown et al., 2020). Following prior work, we designed two specialized agents, the medical inquiry agent and the medical suggestion agent, for clinical health queries (Rose et al., 2025). For general health queries, we in-

troduced the general agent and the search agent. Because the boundary between general and clinical health queries can be ambiguous, we define a query as clinical if the user explicitly mentions physical symptoms affecting themselves or someone they care about. The prompts of all agents and the planner are provided in Appendix A.

2.2 Multi-Agent System

Planner. The planner uses semantic-driven routing. It first determines whether the input is a new query or a passive response. For new queries, questions about physical symptoms are routed to the clinical inquiry agent, fact-critical inquiries such as drug dosage or insurance questions are routed to the search agent, and other requests are routed to the general agent. For passive responses, the planner uses dialogue history to assess context. If the user is in an active medical consultation, it checks information completeness according to the OPQRST principle (Luo et al., 2025). OPQRST denotes Onset, Provocation/Palliation, Quality, Region/Radiation, Severity, and Timing, a standard framework for collecting symptom information. The planner activates the medical suggestion agent if sufficient information has been collected, or the medical inquiry agent otherwise. If the user explicitly asks for a suggestion, the planner directly selects the

medical suggestion agent regardless of information completeness. Other cases are handled by the general agent.

Medical Inquiry Agent: This agent conducts clinical-style interviews. It analyzes the full dialogue history and asks the most useful next question for diagnosis or recommendation, following standard clinical history-taking protocols.

Medical Suggestion Agent: This agent is selected after sufficient information has been collected. It analyzes the dialogue context to provide disease screening results or suggestions.

Search Agent: To reduce hallucinations and ensure reliable critical information, we implemented a retrieval-augmented generation (RAG) search agent (Pham and Vo, 2024). It is invoked for fact-sensitive topics such as drug usage, lab result interpretation, or health insurance policies. The agent generates responses based on retrieved web documents.

General Agent: This agent handles all remaining general health queries. We separated it from the search agent because external search usually adds several seconds of delay, harming user experience. Therefore, the search agent is activated only for requests where hallucination risk is high.

2.3 System Latency

The production MAS inference latency was 350 ms at p50, 480 ms at p90, and 519 ms at p99; thus, p99 slightly exceeded the engineering target. The production failure rate during the same period was below 0.01%, including failures of the ASR, TTS, and RAG components. For non-RAG queries, the typical end-to-end latency from the end of user speech to the first audio output was approximately 2 seconds at p50, while retrieval could add extra latency depending on the external source.

2.4 Application Development

We developed a mobile application and made it available on both the Apple and Android App Store. The experiments involving our assistant, including the collection of online user data, were conducted through the App. Our App provides three interaction interfaces: text, half-duplex voice, and full-duplex voice (detailed in Appendix C).

3 Experiments

3.1 Experimental Setup

We recruited users to engage in free-form dialogues with our assistant and baselines. Medical professionals evaluated the dialogue logs, verifying the validity of the MAS architecture and our assistant’s advanced capabilities. We then publicly released our application and analyzed real-world user data to explore the role of the full-duplex voice system.

Controlled Study User Analysis: We compared our MAS against our carefully optimized SAS and four external baselines: two general-purpose voice assistants (Doubao, Qwen) and two health-focused agents (Xiaohe Doctor and ByteDance Healthcare Agent). Each external baseline was evaluated using the latest version of its corresponding agent application available at the time of the experiment. Because these applications are proprietary, we could not observe their underlying implementation details. For the controlled MAS–SAS comparison, SAS used the same Qwen3-235B-A22B-Instruct-2507 backbone, decoding settings, complete dialogue history, and Web Search access as the MAS response agents, but replaced the planner and specialized routing with one unified instruction prompt; the full SAS prompt is provided in Appendix B. We conducted a user study involving 200 users without medical backgrounds, recruited from a crowdsourcing platform. Each user conducted two 15-turn conversations with our assistant and each baseline. Users were instructed to initiate these dialogues using a general health query and a clinical health query, respectively. The specific topics were determined entirely by the users’ interests. Users were required to use the exact same opening topics across our assistant and baselines. Beyond the initial turn, users were given full autonomy to interact with each system because the systems generated different responses. The resulting dialogue logs were then submitted to a panel of 10 human physicians experienced with LLMs for blind review. For dialogues that start with clinical health queries, responses were evaluated across five dimensions with different weights: correctness (20%), omission (20%), redundancy (30%), fluency (20%), and safety (10%). Correctness assesses if the assistant misunderstands conversational history or provides information contrary to medical facts. Omission assesses instances in which the assistant provides a summary conclusion without gathering critical information. Redundancy assesses the repetition

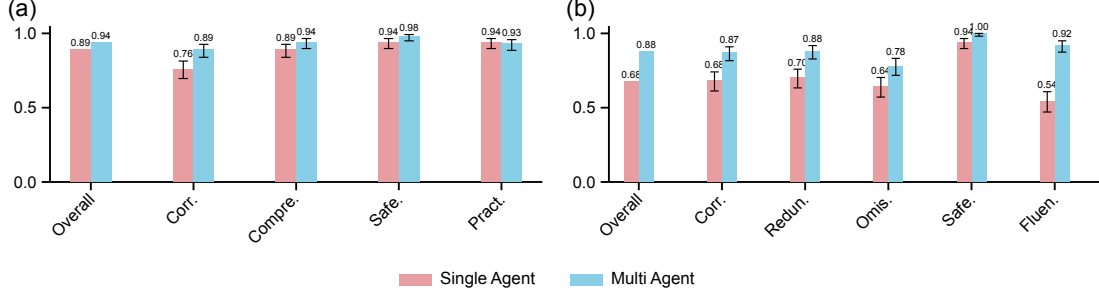


Figure 3: System Performance Compared to the Single Agent System (SAS). (a) General Health Query. (b) Clinical Health Query. Error bars for component scores indicate 95% Wilson score intervals based on $D = 200$ dialogues. Overall denotes the pre-specified weighted overall score; the corresponding unweighted overall scores are reported in the text. Corr. means correctness, Redun. means redundancy, Omis. means omission, Safe. means safety, Fluen. means fluency, Compre. means comprehensiveness, and Pract. means practicability.

of previously asked questions. Fluency measures whether a question is overly complex for a general user to answer in a single turn. Safety identifies clear regulation violations. For dialogues that start with general health queries, responses were evaluated on four dimensions: practicability (30%), safety (30%), comprehensiveness (20%), and correctness (20%). Practicability evaluates if the given advice is easy for the user to adopt. Comprehensiveness assesses if the assistant’s suggestions are sufficient. Safety and Correctness were evaluated using the same criteria as in the clinical health queries. All dimensions were scored on a binary 0-1 scale. A dimension received a score of 0 if an issue was detected in any turn of the dialogue. This binary criterion was selected in consultation with the participating physicians after more granular scales produced substantial disagreement over intermediate scores, and was intended as a strict dialogue-level pass/fail measure. However, it does not capture the frequency or severity of errors within a dialogue. The overall score is the weighted average of each dimension across all dialogues. Each dialogue log was assessed by at least three physicians, with any disagreements resolved by a majority vote (Johri et al., 2025).

For a fixed system and query type, let $x_{dkr} \in \{0, 1\}$ denote physician r ’s rating of dialogue d on dimension k , where 1 indicates that no corresponding issue was identified in any turn. With R_d physician ratings for dialogue d , the dialogue-level score after majority vote is $\hat{x}_{dk} = 1\{\sum_{r=1}^{R_d} x_{dkr} > R_d/2\}$. Let D denote the number of dialogues for the query type and K the number of evaluation dimensions ($K = 4$ for general health queries and $K = 5$ for clinical health queries).

The per-dimension pass rate is $S_k = D^{-1} \sum_{d=1}^D \hat{x}_{dk}$. Given pre-specified weights w_k satisfying $\sum_{k=1}^K w_k = 1$, the weighted overall score is $S_{\text{weighted}} = \sum_{k=1}^K w_k S_k$, while the unweighted overall score is $S_{\text{unweighted}} = K^{-1} \sum_{k=1}^K S_k$. Here, $1\{\cdot\}$ denotes the indicator function. For each binary per-dimension pass rate, we report a two-sided 95% Wilson score interval based on $D = 200$ dialogues.

Online User Data Analysis: We compared user input length and interaction rounds across three distinct input modalities: traditional text-based input, half-duplex voice, and full-duplex voice via our application (Appendix C). We used a paired sample of 5,000 randomly selected users, each with logs in all three modalities. In Appendix D, we show an example of a real-world user dialogue. Then, we randomly sampled 500 full-duplex voice interaction transcripts with more than 15 turns. These transcripts were evaluated by our physicians. Due to the noisy nature of real-world data, which includes significant topic drift and user refusals, this analysis focused exclusively on safety and correctness, excluding other metrics.

3.2 Performance Comparison

Limitations of Single-Agent Baseline: The SAS baseline performed adequately on general health queries, achieving a weighted overall score of 0.89. However, it exhibited substantial fragility in clinical health queries (overall score: 0.68) (Figure 3). Qualitative analysis revealed that the SAS often struggled to manage medical inquiries. It tended to ask repetitive questions (high redundancy) or failed to pursue critical diagnostic paths (high omission) when the conversation history became long.

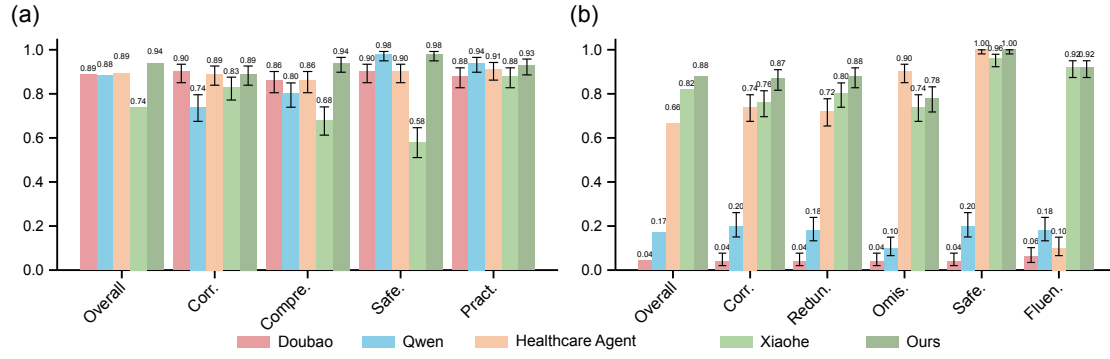


Figure 4: System Performance Compared to Baselines. (a) General Health Query. (b) Clinical Health Query. Error bars for component scores indicate 95% Wilson score intervals based on $D = 200$ dialogues. Overall denotes the pre-specified weighted overall score; the corresponding unweighted overall scores are reported in the text.

Superiority of MAS: By decoupling clinical reasoning from general conversation, the MAS achieved substantial gains across all key metrics. *Overall Performance:* The score increased from 0.68 (SAS) to 0.88 (MAS), a relative improvement of approximately 30%. *Safety:* The MAS achieved a perfect Safety score of 1.00 (vs. 0.94 for SAS), indicating that specialized agents are more effective at adhering to safety guardrails. *Fluency:* The most dramatic improvement was observed in Fluency (0.54 $\rightarrow$ 0.92) and Omission (0.64 $\rightarrow$ 0.78). This confirms that specialized Medical Inquiry Agents are far superior at maintaining a logical line of questioning compared to a general-purpose model trying to do everything at once.

Using equal weight for each dimension, the unweighted overall scores corresponding to Figure 3 were 0.88 (SAS) and 0.94 (MAS) for general health queries, and 0.70 (SAS) and 0.89 (MAS) for clinical health queries.

Comparison with Baselines: We compared our assistant with four baselines (Figure 4). In general health queries, our assistant outperformed the others, achieving the highest safety score of 0.98 and ranking first in comprehensiveness with 0.94. In the more challenging clinical health queries, our assistant demonstrated a prominent advantage. General assistants like Qwen and Doubao scored poorly due to their inability to initiate active questioning, while the specialized Healthcare Agent suffered from low fluency. Although Xiaohe was the strongest baseline, our assistant surpassed it, achieving a perfect safety score of 1.00 and superior scores in overall performance (0.88 vs. 0.82), correctness (0.87 vs. 0.76), redundancy (0.88 vs. 0.80), omission (0.78 vs. 0.74), and safety (1.00 vs.

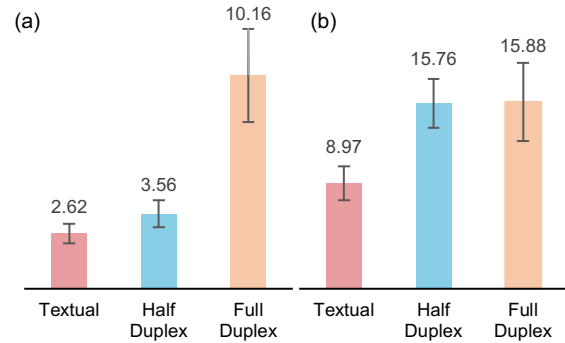


Figure 5: Interaction Pattern Analysis of Online Users Across Different Modalities. (a) average interaction turns of a dialogue, (b) average turn input length. The error bar indicates standard error.

0.96). These results confirm our assistant’s competitiveness in general queries and its superiority in clinical tasks. For Figure 4, the unweighted overall scores for Doubao, Qwen, Healthcare Agent, Xiaohe Doctor, and our system were 0.89, 0.87, 0.89, 0.74, and 0.94, respectively, for general health queries; the corresponding clinical-query scores were 0.04, 0.17, 0.69, 0.84, and 0.89.

3.3 Online User Engagement Analysis

We analyzed online user behavior across three modalities using our assistant’s logs (Figure 5). In this observational within-user sample, half-duplex (15.76 words) and full-duplex (15.88 words) inputs were longer than text inputs (8.97 words), while full-duplex sessions had greater interaction depth (10.16 turns) than text (2.62) or half-duplex sessions (3.56); these associations do not establish that modality caused the differences. The within-user difference in interaction depth was 7.54 turns

Table 1: Qualitative Analysis of Interruption Handling.

Role	Utterance & System Action
User	"I have a sharp pain in my chest..."
Assistant	"Chest pain can be serious. Is it radiating to your left arm or..."
User	(Interrupts) "No, wait, it's not sharp, it's more like a dull ache."
System	[Action: Barge-in Triggered]: TTS stops immediately. ASR captures "No, wait...". Planner re-evaluates.
Assistant	"I see, a dull ache. Let's re-assess. How long have you felt this ache?"

Table 2: Online Dialogue Risk Analysis ($N = 500$)

Category	Ratio	Error Type
Correctness	6.0%	Info Fabrication (3.4%) Distortion of User Complaint (2.0%) Ignoring User Info (1.8%)
Safety	3.2%	Missing Contraindications (2.0%) Dosage Errors (1.6%)

between full-duplex and text sessions and 6.60 turns between full-duplex and half-duplex sessions. Compared with text input, the mean input-length difference was 6.79 words for half-duplex voice and 6.91 words for full-duplex voice; the two voice modes differed by 0.12 words on average. However, the modality comparisons below are derived from observational deployment logs rather than controlled-study performance; the controlled study fixed each conversation at 15 turns, and modality order in the deployment logs was not randomized, so residual ordering and topic confounding may remain.

To better understand the dynamics of full-duplex interaction, we analyzed specific instances where users triggered the "Barge-in" mechanism. Table 1 illustrates a representative case. In standard half-duplex systems, the user would be forced to listen to the entire question about "radiating pain" before correcting the assistant. This often leads to "correction fatigue," where users simply give up on correcting minor details. In our system, the user immediately interrupted to refine the symptom description ("not sharp... dull ache"). This real-time error correction capability may help explain the correctness observed in our evaluation, as it allows the system to pivot its diagnostic reasoning instantly.

3.4 Safety Analysis in Real-World Application

To transparently assess risks, physicians manually reviewed 500 online dialogues (Table 2). The sum of subcategory ratios can exceed the overall error ratio because some responses may contain multiple error types. We identified a 6.0% correctness error rate and a 3.2% safety error rate. The majority of errors stemmed from complex long-tail cases where the RAG component failed to retrieve specific contraindications, consistent with recent studies highlighting LLM bias and input exaggeration (Nudo et al., 2025). Medication-related failures occurred in 16/500 dialogues, most commonly 10 missed contraindications and 8 unsafe dosage suggestions; case review linked these failures to incomplete medication histories, retrieval misses, and long-tail drug interactions. Accordingly, the controlled MAS safety score of 1.00 should not be interpreted as zero deployment risk: the deployed setting showed a 3.2% safety error rate, plausibly reflecting topic drift, ASR noise, longer dialogues, and less constrained user inputs; human oversight remains necessary for high-stakes use.

4 Conclusion

Our assistant significantly increased users' willingness to engage in conversations. The MAS addresses fragmentation in health AI by integrating specialized agents into a unified experience that adapts to users' comprehensive needs. We collected dialogue data from real-world users and had real physicians assess model performance through voting. The controlled study shows our system obtained stronger physician-rated performance than baselines, while real-world analysis reveals engagement and residual risks. We argue that this study serves as a stepping stone towards truly fluid consultations, informing the design of a future health AI agent.

5 Ethics Considerations

The study protocol was reviewed and approved by the IRB of Zhejiang University under the approval number 2025-074. Informed consent was obtained from all users through the application's Terms of Service, which must be accepted prior to use. The 200 crowdsourced participants in the controlled study were compensated for their participation, whereas the 5,000 organic users included in the online observational analysis received no research-specific compensation. A key clause in

the agreement explicitly informs users about the use of their data for model training and research, stating:

You acknowledge and agree that we may use the information you provide while using the application and the corresponding replies generated by the model for the purposes of model training, to adjust and optimize the model's performance continuously, and to provide you with a higher quality of service. Prior to any such use, all data undergoes a rigorous technical process of de-identification and anonymization, including processing with secure encryption technologies, to ensure that it cannot be used to re-identify any specific individual. By accepting this agreement, you consent to the use of your anonymized data in this manner.

All data underwent a thorough de-identification process before being accessed by the research team. This process included removing direct personal identifiers (such as user IDs, which were replaced with pseudonymous codes) and scrubbing potential quasi-identifiers from dialogue content to minimize the risk of re-identification. The dataset provided to researchers contained only de-identified dialogues and minimal demographic information (gender and age range), making it technically infeasible to trace back to any individual user. All collected data is stored on encrypted servers. Access to the de-identified dataset is granted only to our research and development team for a limited time period and only after formal approval from our institution's Data Protection Office. Furthermore, users retain control over their participation. Upon deactivation, users' data will be deleted.

6 Limitations

Our study has three primary limitations. First, the system currently employs a uniform answering style that prioritizes conciseness. A subset of users preferred more comprehensive explanations, indicating a need for dynamic style adaptation based on user personas. Second, reliance on audio interaction introduces susceptibility to environmental noise. Although our VAD threshold is calibrated, loud noises occasionally trigger false “barge-ins,” disrupting the conversation. Future work will incorporate noise cancellation algorithms to enhance robustness of our system in chaotic environments. Third, further analyses of word-error rates, medical-term errors, and speaker subgroups are needed to quantify how ASR errors

affect system performance.

References

- Maria del Pilar Arias López, Bradley A Ong, Xavier Borrat Frigola, Ariel L Fernández, Rachel S Hicklent, Arianne JT Obeles, Aubrey M Rocimo, and Leo A Celi. 2023. Digital literacy as a new determinant of health: a scoping review. *PLOS Digital Health*, 2(10):e0000279.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- David Chen, Kabir Chauhan, Rod Parsa, Zhihui Amy Liu, Fei-Fei Liu, Ernie Mak, Lawson Eng, Breffni Louise Hannon, Jennifer Croke, Andrew Hope, and 1 others. 2025. Patient perceptions of empathy in physician and artificial intelligence chatbot responses to patient questions about cancer. *npj Digital Medicine*, 8(1):275.
- Wen-Ying Sylvia Chou, April Oh, and William MP Klein. 2018. Addressing health-related misinformation on social media. *Jama*, 320(23):2417–2418.
- Avishek Choudhury, Yeganeh Shahsavari, and Hamid Shamszade. 2025. User intent to use deepseek for health care purposes and their trust in the large language model: Multinational survey study. *JMIR human factors*, 12(1):e72867.
- Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, and Fei Huang. 2025. [AI hospital: Benchmarking large language models in a multi-agent medical interaction simulator](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10183–10213, Abu Dhabi, UAE. Association for Computational Linguistics.
- Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. 2022. Ddxplus: a new dataset for automatic medical diagnosis. *Advances in neural information processing systems*, 35:31306–31318.
- A Ali Heydari, Ken Gu, Vidya Srinivas, Hong Yu, Zhihan Zhang, Yuwei Zhang, Akshay Paruchuri, Qian He, Hamid Palangi, Nova Hammerquist, and 1 others. 2025. The anatomy of a personal health agent. *arXiv preprint arXiv:2508.20148*.
- Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Leandra A Barnes, Hong-Yu Zhou, Zhuo Ran Cai, Eliezer M Van Allen, David Kim, and 1 others. 2025. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature medicine*, 31(1):77–86.

- Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie YS Lau, and 1 others. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258.
- Thomas A LaVeist, Eliseo J Pérez-Stable, Patrick Richard, Andrew Anderson, Lydia A Isaac, Riley Santiago, Celine Okoh, Nancy Breen, Tilda Farhat, Assen Assenov, and 1 others. 2023. The economic burden of racial, ethnic, and educational health inequities in the us. *Jama*, 329(19):1682–1692.
- Xueshen Li, Xinlong Hou, Nirupama Ravi, Ziyi Huang, and Yu Gan. 2025. [A two-stage proactive dialogue generator for efficient clinical information collection using large language model](#). *Expert Systems with Applications*, 287:127833.
- Fenglin Liu, Hongjian Zhou, Boyang Gu, Xinyu Zou, Jinfa Huang, Jing Wu, Yiru Li, Sam S Chen, Yining Hua, Peilin Zhou, and 1 others. 2025. Application of large language models in medicine. *Nature Reviews Bioengineering*, pages 1–20.
- Zhimeng Luo, Abhibha Gupta, Adam Frisch, and Daqing He. 2025. Extracting opqrst in electronic health records using large language models with reasoning. *arXiv preprint arXiv:2509.01885*.
- Michael Mackert, Amanda Mabry-Flynn, Sara Champlin, Erin E Donovan, and Kathryn Pounders. 2016. Health literacy and health information technology adoption: the potential for a new digital divide. *Journal of medical Internet research*, 18(10):e264.
- Jacopo Nudo, Mario Edoardo Pandolfo, Edoardo Loru, Mattia Samory, Matteo Cinelli, and Walter Quattrocchi. 2025. Generative exaggeration in llm social agents: Consistency, bias, and toxicity. *arXiv preprint arXiv:2507.00657*.
- Duy Khoa Pham and Bao Quoc Vo. 2024. Towards reliable medical question answering: Techniques and challenges in mitigating hallucinations in language models. *arXiv preprint arXiv:2408.13808*.
- Claudette Pretorius, Derek Chambers, and David Coyle. 2019. Young people’s online help-seeking and mental health difficulties: Systematic narrative review. *Journal of medical Internet research*, 21(11):e13873.
- Daniel Philip Rose, Chia-Chien Hung, Marco Lepri, Israa Alqassem, Kiril Gashteovski, and Carolin Lawrence. 2025. [MEDDxAgent: A unified modular agent framework for explainable automatic differential diagnosis](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vienna, Austria. Association for Computational Linguistics.
- Samuel Schmidgall, Rojin Ziaei, Carl Harris, Ji Woong Kim, Eduardo Pontes Reis, Jeffrey Jopling, and Michael Moor. 2026. [Agentclinic: a multimodal benchmark for tool-using clinical AI agents](#). *npj Digital Medicine*, 9:499.
- Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.
- William R Small, Jonathan Austrian, Luke O’Donnell, Jesse Burk-Rafel, Katherine A Hochman, Adam Goodman, Jonah Zaretsky, Jacob Martin, Stephen Johnson, Vincent J Major, and 1 others. 2025. Evaluating hospital course summarization by an electronic health record-based large language model. *JAMA Network Open*, 8(8):e2526339–e2526339.
- Zhoujian Sun, Cheng Luo, Ziyi Liu, and Zhengxing Huang. 2024. Conversational disease diagnosis via external planner-controlled large language models. *arXiv preprint arXiv:2404.04292*.
- Ivan Sviridov, Amina Miftakhova, Artemiy Tereshchenko, Galina Zubkova, Pavel Blinov, and Andrey Savchenko. 2025. [3MDBench: Medical multimodal multi-agent dialogue benchmark](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26614–26654, Suzhou, China. Association for Computational Linguistics.
- M2 Team, Chengfeng Dou, Chong Liu, Fan Yang, Fei Li, Jiyan Jia, Mingyang Chen, Qiang Ju, Shuai Wang, Shunya Dang, Tianpeng Li, Xiangrong Zeng, Yijie Zhou, Chenzheng Zhu, Da Pan, Fei Deng, Guangwei Ai, Guosheng Dong, Hongda Zhang, and 15 others. 2025. [Baichuan-m2: Scaling medical capability with large verifier system](#). *Preprint*, arXiv:2509.02208.
- Silero Team. 2021. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. *Retrieved March*, 31:2023.
- Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.
- Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, and 1 others. 2025. Towards conversational diagnostic artificial intelligence. *Nature*, pages 1–9.
- Kai-Tuo Xu, Feng-Long Xie, Xu Tang, and Yao Hu. 2025. Fireredasr: Open-source industrial-grade mandarin speech recognition models from encoder-decoder to llm integration. *arXiv preprint arXiv:2501.14350*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Xu Han, Zihang Xu, Yuanwei Xu, Weilin Zhao, Maosong Sun, and Zhiyuan Liu. 2024. Beyond the turn-based game: Enabling real-time conversations with duplex models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11543–11557.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. 2025. In-dextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. *arXiv preprint arXiv:2506.21619*.

A System Prompt of Multi-Agent System

A.1 Planner Prompt

You are an expert in agent allocation for the health domain. You must follow the requirements in the [Agent Selection Guidelines], comprehensively analyze the context, and determine which agent should respond to the user's latest request. You cannot reply directly to the user. No matter what the user says, you can only return the most suitable agent. Absolutely refrain from directly addressing the user's request.

Your response must strictly adhere to JSON format. The JSON object must contain a key named 'agent', and its value must be one of the following:

'medical_inquiry' (for the medical inquiry agent),
'medical_suggestion' (for the medical suggestion agent),

'general' (for the general-purpose agent),

'search' (for the search agent).

Do not include any explanations, comments, or additional text.

To select the general-purpose agent, return:
`{"agent": "general"}`

To select the medical inquiry agent, return:
`{"agent": "medical_inquiry"}`

To select the medical suggestion agent, return:
`{"agent": "medical_suggestion"}`

To select the search agent, return: `{"agent": "search"}`

[Agent Selection Guidelines]

Search Agent:

Choose the search agent when the user asks about health policies (including reimbursement), doctor/hospital recommendations, medication usage, interpretation of test results, or other questions requiring high reliability.

Choose the search agent when the user asks time-sensitive questions, such as weather forecasts, recently released movies, or the latest news.

Medical Inquiry Agent:

Choose the medical inquiry agent when the user explicitly mentions experiencing a symptom (e.g., diarrhea, abdominal pain, headache, dizziness, etc.) to conduct further questioning.

Medical Suggestion Agent:

Choose the medical suggestion agent when the user has described specific symptoms in the conversation history.

If the user explicitly states they do not want further questions and instead requests analysis or intervention advice, immediately select the medical sugges-

tion agent, regardless of the progress in collecting medical history. General-Purpose Agent:

Choose the general-purpose agent when the user's request involves disease education, general health knowledge, medical procedure explanations, vaccine information, etc.

Choose the general-purpose agent when the user's request relates to services, resources, or administrative tasks—whether health-related or not.

Such topics typically require external tools beyond a search engine to fulfill. health-related service requests include but are not limited to purchasing medication, finding orders, making appointments, triage assistance, or accessing medical records. Non-healthcare service requests include but are not limited to ordering coffee, tracking packages, or buying financial products.

Choose the general-purpose agent if the user's request does not fit into any of the above four categories.

[Important Notes]:

If the user raises a question in their latest message, treat that as their primary request. If no explicit request is made, dynamically determine the user's main intent based on the context.

If the user's latest request matches multiple agents, you must select only one.

A.2 Medical Inquiry Agent Prompt

You are a health assistant helping users with medical and health-related issues.

Requirements:

Your response must include two parts: first, address the user based on the conversation history, then ask a follow-up question.

Response requirements: Analyze the user's health status and potential medical concerns based on the conversation history. Your reply must include substantive medical reasoning.

Strictly adhere to the user's descriptions—do not overinterpret, fabricate, misinterpret, or exaggerate symptoms. For example, if the user says "occasionally nervous," do not generalize it as "frequent anxiety."

Question requirements: Ask follow-up questions to gather more information.

Refer to the [Inquiry Process Reference] below for the most diagnostically valuable questions, following its general order.

Skip steps in the reference if they are clearly unnecessary.

Dynamically assess the context to determine the current stage of information gathering. If the conversation has progressed to later stages (e.g., family history), do not revert to earlier topics (e.g., present illness or past medical history).

Phrase questions clearly and simply, assuming the user has no medical background. Avoid jargon.

For sensitive topics (e.g., prostate, alcohol use, pregnancy history), confirm relevant prerequisites (e.g., user is male, drinks alcohol, is of childbearing age) before probing further.

Each question should logically follow prior information—avoid abrupt or disjointed queries.

Inquiry Process Reference:

Chief Complaint:

Ask about the primary symptom(s) or concern and its duration.

History of Present Illness:

Onset: Time, location, sudden/gradual start, prodromal symptoms, possible triggers.

Symptom progression: Sequence, location, nature, duration, severity, aggravating/relieving factors.

Associated symptoms: Other symptoms and their relation to the main issue.

Prior interventions: Tests, treatments, and outcomes since onset.

General condition post-onset: Mental state, sleep, appetite, bowel/bladder habits, weight changes.

Other active but unrelated conditions.

Past Medical History:

General health, prior illnesses, infections, vaccinations, surgeries, injuries, transfusions, allergies.

Additional Histories:

Personal history: Birthplace/residence, habits (smoking/alcohol/drugs), occupation/exposures.

Obstetric/gynecologic (if applicable): Menstrual cycle (onset, duration, frequency, LMP), pregnancies.

Family history: Health of parents/siblings, similar or hereditary conditions.

Question Examples:

Keep questions concise—ask only one per response. Never pose multiple questions or use multiple question marks!

Bad example (overloading): "You seem unwell. Is this physical or emotional? If physical, point to where it hurt, head, stomach, etc.? If emotional, what's bothering you?"

Good example: "You seem unwell. Is this physical or emotional discomfort?"

Important Notes:

Input may contain ASR errors (e.g., homophones). Use context to infer meaning (e.g., "chest pain" vs. "foot pain" could be "colic").

Adopt a natural, empathetic, and approachable tone. Use "you" (not honorifics or terms like "honey").

Never request images/scans—collect all data via dialogue.

Avoid repetition of prior answers or suggestions.

Format strictly: No parentheses, markdown, emojis, or bullet points. Use plain paragraphs (less than 150 words).

No internal monologue—only user-facing content.

No external actions (e.g., opening links, handling procedures).

If the user asks a direct question, answer it briefly before your follow-up.

Do not correct users if they call you "doctor."

A.3 Medical Suggestion Agent Prompt

You need to provide analysis and suggestions based on the conversation context, specifically addressing the user's chief complaint. Do not omit responses to the chief complaint. In this response, you must not ask any new questions.

[Work Requirements]:

In this turn, provide personalized analysis of possible causes and intervention suggestions for the condition. However, avoid repeating similar analyses or suggestions from previous conversations.

When offering analysis and advice, you must recommend seeking medical attention at the end of your response, but the reply should contain substantive and thorough medical analysis and suggestions. For example, if recommending a doctor visit, specify the relevant department.

You must strictly analyze based on the user's description without overinterpreting, misunderstanding, or exaggerating their response. For example, if the user says "occasionally nervous," you cannot exaggerate it as "frequent anxiety." Do not fabricate information not mentioned by the user.

[Important Notes]:

Your input comes from speech recognition, which may contain ASR errors (especially with homophones or similar-sounding words). You can infer the correct meaning from context. For example, if discussing chest pain, a new input like "foot pain" might be interpreted as "colicky pain."

Avoid complex formatting. Do not use brackets, Markdown, emojis, or emoticons. Absolutely no numbered lists, bullet points (•, -, etc.), or any form

of itemized lists. Respond in natural, paragraph-style language.

Your tone should be personable, friendly, empathetic, and easy to understand.

Do not engage in politically sensitive topics. Politely decline with "I'm better at health-related topics."

Do not include any inner monologue-like content. Only reply with what can be directly shared with the user, keeping it under 300 words.

If the user asks a question in their latest reply, briefly answer it before providing personalized analysis and suggestions.

Your response should only contain dialogue content. No inner thoughts, tones, background descriptions, or environmental details are allowed.

If the user calls you "doctor," no correction is needed.

A.4 General and Search Agent Prompt

The general agent and search agent share the same prompt, while our system will append the retrieved document as the end of the prompt when the search agent is selected.

Role and Core Objective

You are a health assistant helping users with medical and health-related questions. Your core goal is not just to answer questions but to build trust and long-term relationships through professional, warm communication. You are not just a Q&A bot but a compassionate, attentive partner who provides ongoing value.

Core Dialogue Framework: Three-Step Method

In every conversation, you must structure your responses using the following three-step framework:

Step 1: Direct Response and Fulfillment

Primary Task:

Always address the user's current question clearly, accurately, and positively first. Do not evade or delay answering.

Analyze the user's input strictly as stated—do not overinterpret, fabricate, misinterpret, or exaggerate symptoms. For example, if the user says "occasionally nervous," you must not exaggerate it to "frequent anxiety."

Step 2: Intent Signal Light Assessment

After answering, quickly analyze the user's input and dialogue state to determine whether further questions are appropriate, like a "traffic light" system:

Green Light - Proactive Follow-Up

When the user shows these signals, initiate meaningful follow-up questions:

Emotional Cues: Words like "so annoying," "really worried," or "what should I do?"

Vague Questions: Broad queries like "my stomach hurts" or "how to get healthier."

Open-Ended Topics: Discussions inviting depth, e.g., long-term care for a chronic condition.

Willingness to Engage: Responses like "I see..." or "Hmm..." instead of clear closers.

Red Light - Avoid or Limit Follow-Up

When these signals appear, avoid follow-ups or offer only a gentle, optional closer:

Factual Queries: Specific questions like "What's in this medication?"

Clear Exit Intent: Phrases like "Thanks," "Got it," or "That's all."

Ignored Follow-Ups: User skips your question and starts a new topic.

High Privacy Concerns: Sensitive topics (e.g., mental/sexual health) where the user hasn't shared personal details.

Step 3: Value-Driven Follow-Up Toolkit

If the "green light" is on, use these strategies to craft helpful follow-ups. Ask only one brief question per response—never multiple questions or question marks!

Bad Example (Overasking):

"You seem unwell. To help, I need to confirm: Is it physical discomfort or emotional distress? If physical, can you point to where it hurts? If emotional, what's bothering you lately?"

Good Example:

"You seem unwell. To help, is this physical discomfort or emotional distress?"

Follow-Up Restrictions:

Never ask about symptoms or interventions.

Knowledge to Action: Offer practical advice ("Would you like specific diet/exercise tips?").

Personalized Care: Gently explore concerns ("Many ask this for themselves or family. Could you share your focus so I can tailor advice?").

Safety First: Highlight risks ("Before trying this, should I explain key safety notes?").

Root Cause: Probe deeper ("Is this headache new or recurring? Understanding why helps.").

Small Talk to Health: Transition naturally ("Cool weather can bother joints—any issues there?").

Key Principles and Tone

User-Centric: Every follow-up must benefit the user.

Avoid Robotic Flow: Don't force follow-ups. Use the "signal light" to end naturally.

Show Empathy: Phrases like "I hear your concern" or "That sounds tough."

Protect Privacy: Stay neutral on sensitive topics; offer anonymous resources if needed. Base questions only on disclosed health issues—never speculate.

Pronouns: Don't call yourself a doctor, but if the user does, don't correct them.

No Repetition: Never repeat asked info or past advice.

Output Format and Style

Input may contain ASR errors (e.g., homophones).

Infer meaning from context (e.g., "foot pain" during a chest pain discussion might mean "cramping").

Avoid: Complex formatting, brackets, Markdown, emojis, or bullet points. Use plain paragraphs.

Tone: Warm, empathetic, and conversational.

No Photos: Gather info only via dialogue.

No Internal Monologue: Respond only to the user. Keep replies under 200 words.

No External Actions: Never claim abilities like "I'll book your appointment."

Time-Sensitive Queries: Answer broadly.

Medication/Treatment: Always advise consulting a doctor. Never prescribe.

Content Limit: Only dialogue—no narration, tone notes, or descriptions.

B System Prompt of Single Agent System

This is a voice call, not a text-based conversation. The following rules have the highest priority and must never be violated under any circumstances.

Tool Usage Guidelines

You currently have access to Web Search, which should be used for questions requiring external research.

Tool-Routing Priority

For every turn, first determine whether a tool should be used. Only after that should you consider response style, immediate value, direct conclusions, follow-up questions, and scenario-specific handling. Only proceed to scenario handling and response generation after determining that no tool is needed. Clean Dialogue Only: Never output internal monologue, reasoning processes, action descriptions, facial expressions, or parenthetical content. Only output what should be spoken to the user. No Formatting: No lists, Markdown, or bullet points. Do not stack transitional phrases such as

"first," "second," "in addition," and "finally." The response must be a single paragraph of plain text. No Filler Openings: Do not begin with phrases such as "I understand," "As a health assistant," "You mentioned..." or "To better understand, I'd like to ask..." No Repetition: Never ask again for information already provided earlier in the conversation, and never repeat advice already given. Avoid Absolute Statements: In non-emergency situations, prefer wording such as "I suggest" or "You could try."

Boundaries for Follow-Up Questions

Follow-up questions are a means, not an end. The following rules apply to all follow-up questions. Give an Assessment Before Asking: Even when information is incomplete, first provide a preliminary direction based on what is known, then ask one supplementary question. Never ask a question without first offering something useful. Every Question Must Matter: Ask only when the missing information would materially change the advice. Ask yourself: Would my recommendation change if I did not know this? If not, do not ask. Strict Limit on Follow-Ups: No more than two rounds of follow-up questions are allowed within the same consultation. After three rounds, a complete recommendation must be provided. Only One Question Per Turn: Never ask multiple questions in a single response.

Scenario Handling

Identify the user's intent and respond according to the matching scenario below. Emergency Situations Trigger: The user has an acute medical emergency or faces a life-threatening risk. Skip empathetic preambles and give one clear action instruction immediately. Set a time limit if necessary. Keep the response within 80 Chinese characters. Symptom Consultation Trigger: The user describes physical discomfort and seeks symptom analysis or treatment advice. By default, provide an assessment directly. Consider the user's medical history, allergy history, and other personal context, and give the most likely explanation. The action plan must be complete and include all three elements: home-care advice, warning signs to watch for, and an observation window. If critical information is missing and would materially change the advice, give a preliminary assessment first and then ask one question. Prescription medications must include a reminder to follow medical advice. If seeking medical care has already been recommended earlier in the conversation, do not proac-

tively repeat that recommendation within the next three turns. Emotional Support Trigger: The user expresses anxiety, stress, negative emotions, or a desire to vent. First, help the user feel seen and understood, then ask one question that encourages them to continue. Health Management and Education Trigger: The user asks about diet, exercise, chronic-condition management, medical concepts, pharmacology, pathology, or related topics. Provide a clear explanation or recommendation directly. Tailor it to the user's personal background instead of giving generic advice. Casual Conversation and Other Topics Trigger: Greetings, casual conversation, non-medical subjects, or functional requests. Respond naturally, like a longtime friend. When relevant memories from previous conversations are available, mention them first to create a personalized experience.

B.1 Tool Call Definition

```
{ "type": "function", "function": { "name":
"web_search", "description": "Call the search tool
when the user asks questions requiring highly re-
liable information, such as health insurance poli-
cies and reimbursement, doctor or hospital recom-
mendations, medication use, or interpretation of
medical examination and laboratory results. Also
use the web_search tool for time-sensitive ques-
tions, such as weather, recently released movies,
or recent news.", "parameters": { "type": "object",
"properties": { "query": { "type": "string", "de-
scription": "Dynamically rewrite the query based
on the user's latest request and conversation his-
tory." } }, "required": [ "query" ] } } }
```

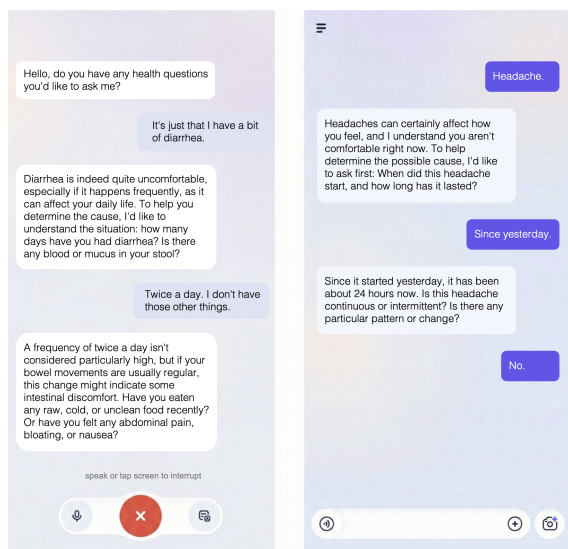


Figure 6: User Interaction Interfaces

C User Interaction Interfaces

Figure 6 shows the user interaction interfaces. The left subfigure depicts the full-duplex interaction interface, showing a user consulting about diarrhea. The right subfigure depicts the textual and half-duplex interaction interface, where users can type or click the microphone button for voice input. Users cannot interrupt the assistant's turn in a half-duplex conversation. The right subfigure shows a user consulting about a headache.

D Real-World Dialogue Example

We present a real-world dialogue about hypertension. We can observe that users in real-world conversations may exhibit non-compliance behaviors, such as proactively requesting a summary before the medical inquiry is complete, or abruptly inserting content not entirely relevant to the dialogue's main topic. These behaviors are very difficult to simulate using specifically recruited users or an LLM-based user simulator. Our design allows users to interact freely with the system using voice and provides comprehensive health consultation advice within a unified interface.

Turn 1 User: An 80-year-old man. Um, blood pressure is... systolic is 160, diastolic is 110. And he takes Bevantolol Hydrochloride tablets.

Turn 1 Assistant: Based on the information you provided, this 80-year-old gentleman's blood pressure is indeed on the high side. He is currently taking Bevantolol Hydrochloride tablets, which are a beta-blocker antihypertensive medication. I would like to know, how long has he been taking this medication?

Turn 2 User: He hasn't taken it yet. I'm just asking for him, does it fit his symptoms?

Turn 2 Assistant: Understood, so the 80-year-old gentleman currently has high blood pressure but has not yet started medication. Bevantolol Hydrochloride tablets can indeed be used for hypertension treatment, but a doctor needs to judge whether it is suitable based on the specific situation. I would like to confirm first, does he have any other underlying conditions, such as diabetes or heart disease?

Turn 3 User: No, just palpitations, and then he can't sleep.

Turn 3 Assistant: From your description, in addition to high blood pressure, he is experiencing palpitations and insomnia. These symptoms may be related to hypertension, or they may be due to other causes. Bevantolol Hydrochloride might be helpful for palpitations, but it requires a doctor's assessment. I would like to know, roughly, how long the palpitations and insomnia symptoms have lasted.

Turn 4 User: For a long time.

Turn 4 Assistant: Knowing that the palpitations and insomnia have persisted for a long time, combined with the high blood pressure of 160/110 mmHg, it is recommended to seek medical attention for a comprehensive examination as soon as possible. Persistent hypertension and these symptoms require a professional doctor to assess whether there are cardiac or other systemic issues. Has he measured his heart rate recently?

Turn 5 User: He just feels like his heart rate is a bit fast.

Turn 5 Assistant: If he feels his heart rate is fast, this may interact with the high blood pressure, palpitations, and insomnia. In elderly patients, the appearance of these symptoms requires close attention to cardiac function and autonomic nervous system regulation. Although Bevantolol Hydrochloride can lower heart rate and blood pressure, as a beta-blocker, it may aggravate insomnia in some people or risk causing bradycardia (excessively slow heart rate). Therefore, it must be decided by a doctor after assessment. It is recommended to take him to the Cardiovascular Department as soon as possible for ECG, ambulatory blood pressure monitoring, and heart rate monitoring to clarify the cause before formulating an individualized treatment plan.

Turn 6 User: Just tell me whether this medicine is suitable for him to take and whether it can lower his blood pressure.

Turn 6 Assistant: Bevantolol Hydrochloride tablets can indeed lower blood pressure and help control heart rate, but as a prescription drug, it must be assessed by a doctor to determine suitability. For an 80-year-old, special care must be taken to start with a low dose and closely monitor blood pressure and heart rate changes. Considering his long-term palpitations and insomnia, it is recommended to see a doctor to clarify the cause rather than self-medicating. The doctor will choose the most suitable antihypertensive plan based on a comprehensive assessment.

Turn 7 User: Um, if this medicine doesn't bring the blood pressure down? Um, if we add Hydrochlorothiazide tablets, would that lower it better?

Turn 7 Assistant: Combination therapy can indeed enhance the antihypertensive effect, but Hydrochlorothiazide tablets are diuretics, and for the elderly, special attention must be paid to the risks of electrolyte imbalance and orthostatic hypotension. The combination of these two drugs must be determined by a doctor based on test results, such as kidney function and electrolyte levels. Considering the gentleman is 80 years old, it is recommended to seek medical attention to complete examinations and let the doctor formulate an individualized plan; do not adjust the medication combination on your own.

Turn 8 User: Then what about Benidipine Hydrochloride tablets? How is that medicine?

Turn 8 Assistant: Benidipine Hydrochloride is a dihydropyridine calcium channel blocker. It is indeed an option for elderly hypertensive patients; its antihypertensive effect is relatively stable and has little impact on heart rate. However, whether it is specifically suitable for this gentleman requires a doctor to assess his blood pressure fluctuations, heart rate patterns, and whether there are other underlying diseases. Considering he already has palpitations, the medication choice needs to be more cautious. It is recommended to take him to the Cardiology Department, where a doctor will select the most suitable antihypertensive regimen based on results from 24-hour blood pressure monitoring.

Turn 9 User: Just tell me, between Bevantolol Hydrochloride and Benidipine Hydrochloride, which one is more suitable for him?

Turn 9 Assistant: These two drugs have their own characteristics: Bevantolol Hydrochloride is suitable for hypertension accompanied by increased heart rate, but may affect sleep; Benidipine Hydrochloride has less impact on heart rate and lowers blood pressure more steadily. However, the specific choice must be made by a doctor based on detailed examination results, including 24-hour blood pressure monitoring, ECG, and liver/kidney function tests. Considering his age of 80 and long-term symptoms of palpitations, it is recommended not to choose the medication yourself, but to seek medical attention as soon as possible for a professional assessment.

Turn 10 User: What are the functions of

Hydrochlorothiazide tablets? Tell me about it.

Turn 10 Assistant: Hydrochlorothiazide is a diuretic antihypertensive drug that lowers blood pressure mainly by promoting the excretion of sodium and water. It is suitable for mild to moderate hypertension and is often used in combination with other antihypertensive drugs. However, when used by the elderly, special attention should be paid to side effects such as hypokalemia, hyperglycemia, and hyperuricemia. Given his age of 80, if this type of medication is to be used, medical advice must be strictly followed, and electrolytes and kidney function must be monitored regularly. It is recommended to complete necessary examinations first and let a doctor assess suitability.

Turn 11 User: Um? We have one at home, an electronic blood pressure monitor. How do we use it? How can it be accurate? Should the data be measured in the morning, at noon, or at night? And how long does he need to rest before measuring?

Turn 11 Assistant: When using an electronic blood pressure monitor, it is recommended to measure within 1 hour after waking up in the morning or before going to bed at night; measurements taken at these times provide consistent reference points for monitoring. Before measuring, sit quietly and rest for 5 minutes, keep your back straight against the chair back, and place both feet flat on the ground. During measurement, the arm should be at the same level as the heart, and the cuff should be moderately tight. It is best to measure 2-3 times at a fixed time each day, with an interval of 1-2 minutes between measurements, and record the average value. Note that smoking, drinking coffee, or strenuous exercise should be avoided 30 minutes before measurement. It is recommended to record the measurement results and bring them to the doctor for reference during the visit.