STELLA: An Open-Source Platform for Auditable Voice-First Conversational Health Coaching Agents

Felix Moser
School of Medicine
University of St. Gallen
St. Gallen, St. Gallen, Switzerland
Institute for Implementation Science in Health Care
University of Zurich
Zurich, Switzerland
felix.moser@unisg.ch

Rasita Vinay
Institute of Biomedical Ethics and History of Medicine
University of Zurich
Zurich, Switzerland
Department of Management, Technology, and Economics
ETH Zurich
Zurich, Switzerland
rasita.vinay@ibme.uzh.ch

Tobias Kowatsch
Institute for Implementation Science in Health Care
University of Zurich
Zurich, Switzerland
School of Medicine
University of St.Gallen
St.Gallen, Switzerland
tobias.kowatsch@uzh.ch

Abstract

Deploying voice agents powered by LLMs creates tension between system auditability and the low latency required for natural conversation, while imposing heavy infrastructure burdens. We present STELLA, an open-source platform enabling the efficient deployment of voice-first agents. As a platform, STELLA provides ready-to-use infrastructure, including plan-based session management, modular expert pipelines, and a developer SDK. This eliminates the need for researchers to build complex audio and safety backends from scratch. To address the latency and auditability tension, STELLA's default agent structurally decouples conversational presence from backend processing. Upon user input, the agent immediately delivers backchannel acknowledgments to maintain natural conversational flow, while safety arbitration and expert routing execute concurrently. A 32-participant field evaluation showed STELLA achieves 1.5-second median latency on commodity scientific hardware and above-median conversational quality in the 50th to 75th percentile ASAQ representative set, demonstrating feasibility for health coaching deployment and motivating future work on clinical applications such as cognitive stimulation therapy.

CCS Concepts

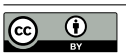
• **Human-centered computing** → **Interactive systems and tools**; *Sound-based input / output*; User studies; • **Computing methodologies** → *Discourse, dialogue and pragmatics*; • **Applied computing** → Health informatics.

Keywords

conversational agents, voice user interfaces, clinical auditability, health coaching, open-source, LLM

ACM Reference Format:

Felix Moser, Rasita Vinay, and Tobias Kowatsch. 2026. STELLA: An Open-Source Platform for Auditable Voice-First Conversational Health Coaching Agents. In *ACM Conversational User Interfaces 2026 (CUI '26), July 21–24, 2026, Bremen, Germany*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3816046.3816293



This work is licensed under a Creative Commons Attribution 4.0 International License.
CUI '26, Bremen, Germany
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2741-2/26/07
https://doi.org/10.1145/3816046.3816293

1 Introduction

Large language models (LLMs) are transforming conversational health coaching through open-ended, adaptive interactions that more closely resemble natural human dialogue [10, 17]. Early evidence links this naturalness to stronger user engagement and higher perceived empathy [20]. When combined with voice interaction, these systems become particularly promising for populations where text-based interfaces create barriers, such as older adults and individuals with cognitive impairments [16, 22]. In the tradition of relational agents [3], we define a *coaching agent* [1] as a conversational system specialized in supporting users toward behavioral or therapeutic goals through structured, plan-based dialogue.

However, deploying voice-based coaching in clinical settings exposes a fundamental tension between *auditability* and *conversational agency*. High-risk AI systems must support automatic record-keeping sufficient for traceability and post-hoc inspection of their behavior, as mandated by frameworks like Article 12 of the EU AI Act [12]. Achieving this requires significant computational overhead, such as logging expert verdicts, arbitration directives, and plan states for every conversational turn (see the supplementary material for an example trace). This overhead occurs precisely when system responsiveness is most critical: human turn-taking naturally operates at gaps of approximately 200 milliseconds [21]. The latency introduced by rigorous safety processing easily disrupts this interactional flow, risking the degradation of the system into a rigid, transactional interface [27]. Currently, commercial platforms sacrifice auditability for speed, while open-source frameworks provide transparent audio plumbing but do not address the session management and safety infrastructure that clinical studies require.

To bridge this gap, we present STELLA (**S**ystem for **T**esting and **E**ngineering **LL**M-based conversational **A**gents), an open-source platform[1] for building auditable, low-latency, voice-first conversational agents. STELLA's core design contribution is the structural decoupling of a conversational agent's *presence* (its immediate responsiveness) from its *reasoning and safety* layers. Upon receiving

[1] Source code, license, deployment guide, evaluation configuration, expert prompts, and conversation plans are available at https://github.com/c4dhi/STELLA.

user input, the agent immediately produces a backchannel acknowledgment to maintain conversational flow. Concurrently, safety arbitration and expert routing execute in the background. By design, this architecture absorbs the computational cost of auditability during the acknowledgment phase, eliminating disruptive pauses. Architecturally, STELLA distinctly separates the underlying *platform* (handling deployment, orchestration, and voice infrastructure) from the *agent* logic (governing conversational behavior). It exposes an SDK for authoring structured conversation plans, allowing researchers to host custom agents or utilize the provided default agent, which is already optimized for parallel safety arbitration.

We evaluate STELLA through a field deployment with 32 participants. Running on a single, modest commodity node (NVIDIA Tesla T4, 8 VCPUs, 32GB RAM), the system's conversational quality ranked within the 50th to 75th percentile of established social agents, as measured by the Artificial Social Agent Questionnaire (ASAQ) [13]. Finally, we discuss the implications of this architecture for clinical deployments, drawing on recent work that delivers Cognitive Stimulation Therapy to individuals with early dementia [35, 36].

2 Related Work

Conversational agents for health behavior change have a long research lineage. Early work by Bickmore et al. [3, 4] demonstrated that relational agents can build therapeutic alliances even within rule-based systems. Modern LLMs have expanded these capabilities. Systems such as GPTCoach [20] implement motivational interviewing through prompted dialogue, while MIcha [28] shows that GPT-4-powered support can match traditional interventions in a randomized trial. However, these systems are primarily text-based, limiting accessibility for populations that benefit from voice interaction [16, 22]. Voice-based alternatives remain underexplored. For example, Talk2Care [39] focuses on information delivery rather than structured therapeutic interaction.

Black Box Agency. Deploying structured, low-latency voice interventions currently forces researchers into two unsatisfactory paradigms. Commercial platforms such as Woebot [14] and Lark [2] deliver evidence-based coaching but are limited to text and proprietary architectures. In contrast, voice-native platforms such as Hume AI [18], Vapi [34], Bland AI [5], and ElevenLabs [11] demonstrate that low latency is achievable in practice. However, these platforms operate as structural black boxes, offering no built-in support for inspectable safety monitoring, structured multi-phase plans, or systematic data extraction. As a result, they provide conversational agency without auditability, limiting their suitability for clinical research and reproducibility.

Fragmented Auditability. Open-source frameworks provide the transparency required in clinical contexts but remain unable to deliver conversational agency by themselves. Orchestration frameworks such as LangChain [6] and AutoGen [38], alongside safety toolkits like NeMo Guardrails [29], are engineered for text-based, asynchronous workflows and do not support real-time turn-taking or backchanneling [23]. Voice-capable frameworks like LiveKit Agents [24] address audio transport but provide no structured conversation plans, safety monitoring, or persistent state management. Consequently, researchers must independently integrate an audio pipeline, a safety arbitration layer, a conversational state machine, and a data backend before any ethics board review.

No existing platform, to our knowledge, jointly supports voice-first interaction, auditable safety architecture, and plan-based session management for structured health interventions (Table 1). STELLA is designed to address this gap. By providing these layers as an integrated, deployable system, STELLA reduces the engineering overhead required to bring auditable voice agents from prototype to field deployment.

3 System Design

STELLA comprises two layers: a *platform* managing sessions and infrastructure, and a configurable *default agent architecture* providing a foundation for structured coaching. Researchers can utilize this default pipeline or implement custom agents via the STELLA Agent SDK.

3.1 STELLA Platform

The platform separates experimental configuration from agent reasoning (Figure 1). Interactions occur within isolated *sessions* powered by WebRTC (LiveKit [25]).

3.1.1 Plan and State Management. Sessions are guided by *conversation plans*, which are hierarchical JSON objects defining states, tasks, and deliverables. Crucially, the platform persistently stores the complete session state, including dialogue history and plan progress. Rather than maintaining context in active memory, the agent dynamically requests its required context (such as state snapshots and recent history) from the platform on each turn. Researchers retain full control over exactly what information is retrieved and injected, ensuring stateless resilience and seamless recoverability.

3.1.2 Shared Voice Infrastructure. STELLA abstracts voice processing into pluggable STT and TTS components. Rather than embedding these engines within individual agents, STELLA deploys them as shared, GPU-accelerated microservices. This centralizes concurrency management and eliminates per-session GPU overhead. The custom SDK also enables researchers to build agents using simple text interfaces while the platform transparently handles gRPC routing and audio streaming.

3.2 Default Agent Architecture

Real-time voice interaction necessitates fast, lightweight LLMs. However, prompting a single model to simultaneously evaluate audio clarity, track goal progression, and monitor user safety frequently degrades reliability, resulting in missed signals or inconsistent behavior [19]. To address this, the default agent decomposes the computational workload into a four-stage pipeline. This structural separation guarantees that decisions remain transparent and auditable, while the inherent latency of these rigorous background checks is masked through immediate backchanneling to preserve perceived responsiveness (Figure 2).

3.2.1 Stage 1: Classification and Backchanneling (S1). An *Input Gate* routes user utterances to relevant experts. Concurrently, a *Bridge Generator* emits a brief acknowledgment (e.g., "I hear you")

Table 1: Requirements for auditable voice coaching agents (motivated in Section 2 preceding paragraphs) and existing solutions that address each. No single system addresses all four; STELLA integrates them by building on LiveKit for voice transport, adopting plan abstractions in the spirit of openCHA, and adding inspectable safety arbitration.

Requirement	Addressed by	What is still missing
Real-time voice interaction	Commercial voice platforms [5, 11, 18, 34]; LiveKit Agents [24]	Commercial: closed source, no inspectable safety. LiveKit: no plan or safety layer.
Inspectable safety routing (EU AI Act Art. 12 [12])	NeMo Guardrails [29], LlamaFirewall [7]	Text pipeline only; no voice integration or dialogue plan abstraction.
Structured intervention delivery (plan-based dialogue)	openCHA [1]; LangChain [6] and AutoGen [38] as orchestration substrate	Text based; not designed for real-time voice or clinical safety routing.
Open-source reproducibility	LangChain, AutoGen, openCHA, LiveKit Agents (as above)	No single system jointly provides voice, audit, and plans as an integrated platform.

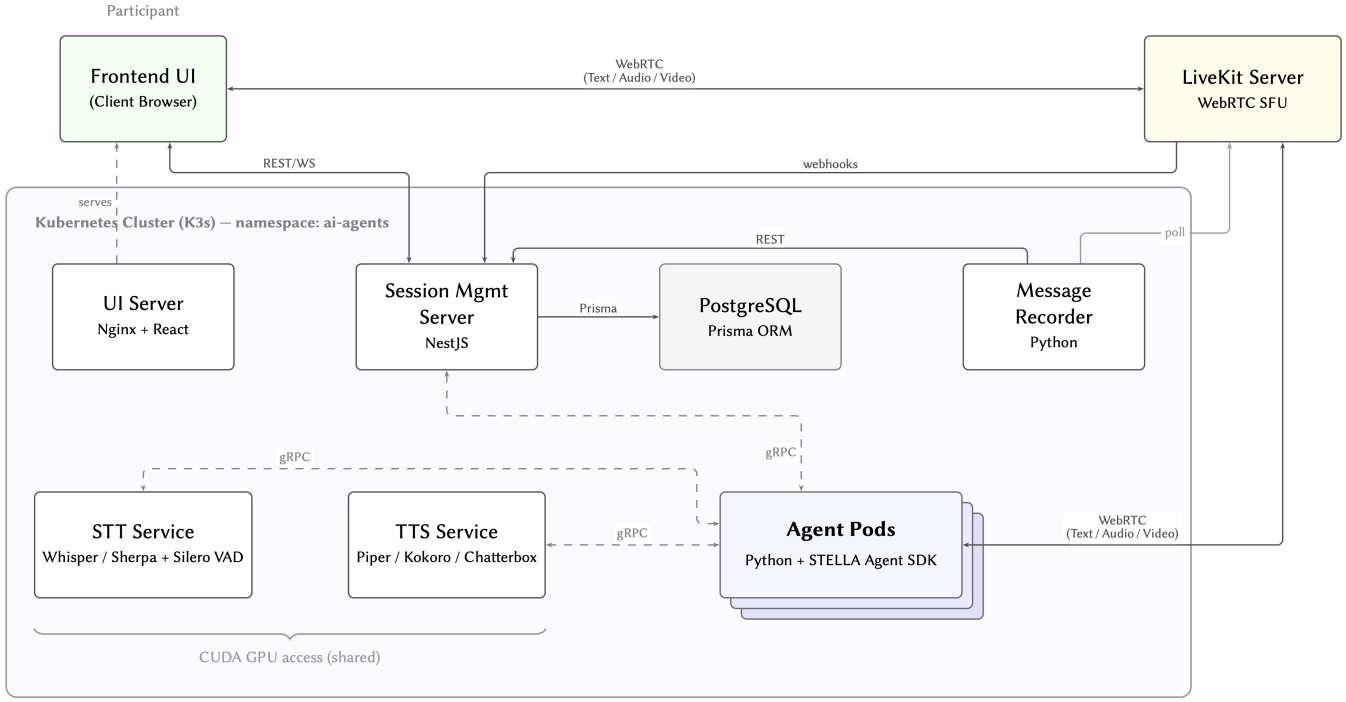


Figure 1: STELLA platform architecture. The session management server orchestrates agent pods, shared voice microservices, and real-time audio transport via LiveKit.

via TTS. This immediate backchannel maintains the conversational floor [31] and masks the latency of subsequent safety audits.

3.2.2 Stage 2: Parallel Expert Pool (S2). To maintain responsiveness while monitoring multiple dimensions, STELLA distributes the analytical workload. Specialized prompts evaluate isolated dimensions (e.g., noise detection, timekeeping, task extraction) concurrently. This decomposition is designed to isolate analytical tasks, aiming to preserve detection reliability without compounding the latency of a single generation step. Researchers can independently tune each expert's context window to balance analytical depth with computational overhead.

3.2.3 Stage 3: Deterministic Arbitration (S3). To promote consistent system behavior, STELLA separates action synthesis from the generative model. Expert verdicts are aggregated and resolved via a strictly rule-based priority hierarchy where systemic signals can override task-level decisions. For instance, if the noise expert flags garbled audio (priority 100), this verdict short-circuits lower-priority task extraction (priority 70) and forces a clarification request. This rule-based substitution is designed to support transparent, auditable actions.

3.2.4 Stage 4: Response Generation (S4). The generator composes the final response using the resolved arbitration directive, plan state, and prior bridge. The final result in text form is streamed to TTS

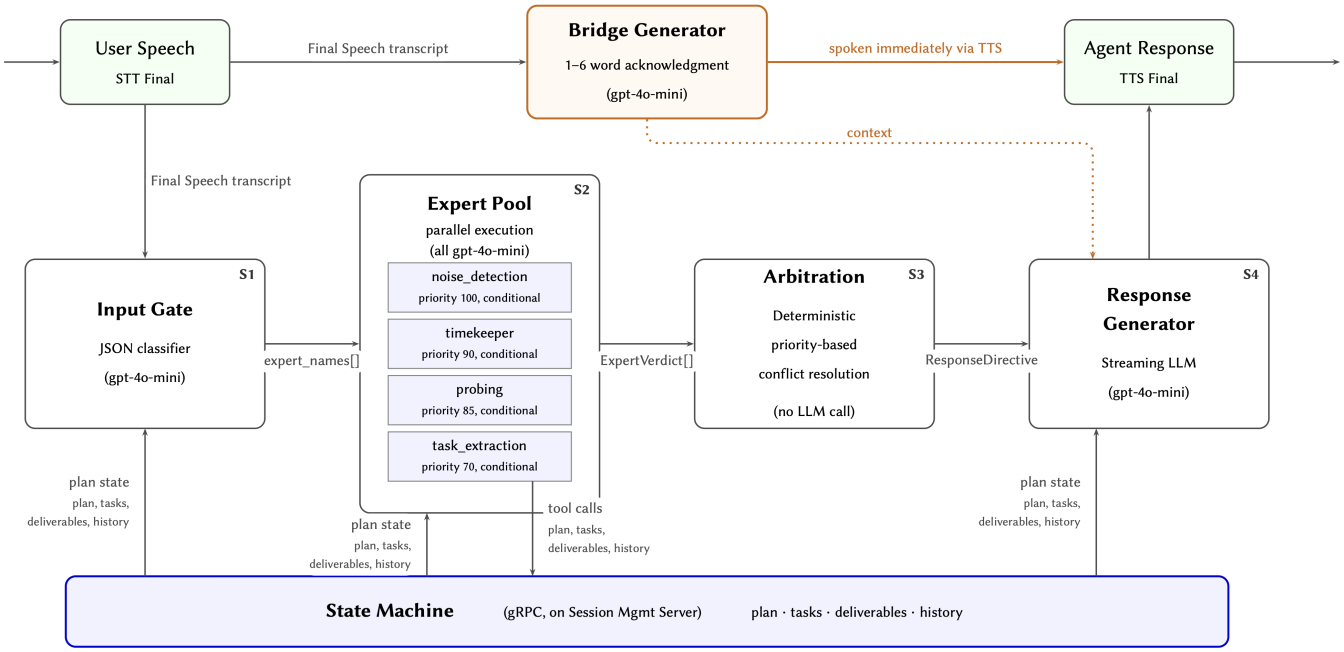


Figure 2: Default agent pipeline as configured for the evaluation. Stages S1–S4 process each user turn from classification through response generation.

sentence-by-sentence. While Stage 1 establishes conversational *presence*, Stage 4 ensures a consistent coaching *identity* [32].

4 Evaluation

To assess conversational quality, we evaluate STELLA's default agent configured for physical activity coaching (four experts; see Section 3) using two constructs from the validated Artificial Social Agent Questionnaire (ASAQ) [13]: *Natural Behaviour* ("the extent to which a user believes that the social agent's behaviour could exist in or be derived from nature") and *Agent's Coherence* ("the extent to which the agent is perceived as being logical and consistent"). We additionally report time-to-first-audio (TTFA), the interval from end of user speech to onset of the backchannel acknowledgment, as an objective measure of system responsiveness. ASAQ scores are contextualized against the ASAQ representative set 2024, a normative dataset of 29 agents (including voice assistants, social robots, and virtual humans) evaluated by 1,066 participants.

Participants and procedure. We recruited participants via Prolific [30]. A pretest with 9 participants (5 Prolific, 4 research team) verified the technical setup. Pretest data were excluded from all analyses. In the main study, 46 participants consented. Of these, 14 did not complete the study: 7 due to technical issues (microphone not detected, session timeout of seven minutes hit, or premature conversation abandonment) and 7 who abandoned before the voice interaction, resulting in 32 complete responses. As a single-system evaluation, no power analysis was required. With $n = 32$, 95% CIs are ±0.61 (NB) and ±0.44 (AC), sufficient to locate the agent within the normative distribution.

The study was administered via LimeSurvey hosted at the last author's university. After informed consent, participants completed demographics and three voice assistant experience items (frequency of use, comfort level, and self-rated experience, each on a 5-point scale). They then followed a step-by-step tutorial and interacted with the agent via voice about exercise habits and barriers (all pipeline stages active, approx. 3–7 min). Afterward, they completed two ASAQ constructs, Natural Behaviour (3 items) and Agent's Coherence (4 reverse-scored items), both on a bipolar scale from −3 to +3, and two open-ended questions. TTFA was logged automatically by STELLA for every conversational turn.

Ethics. As the scope of this study does not fall within the National Human Research Act, formal ethics approval was not required. Informed consent was obtained prior to participation, and participants received monetary compensation via Prolific. Voice input was transcribed in real time on the last author's university infrastructure and never stored as audio. Prompts, including transcribed texts, were sent to OpenAI's GPT-4o mini via a paid API account for response generation.

Analysis. Agent's Coherence items were reverse-scored ($\times -1$). Construct scores were computed as item means. We report M, SD, 95% CI, and Cronbach's α. We contextualize ASAQ scores using Tran et al.'s percentile classification [33] (low: <50th; moderate: 50th–75th; high: 75th–90th).

Results. The 32 participants who completed the study were aged 18–56 ($M = 31.2$, $SD = 8.8$; 15 male, 17 female) and resided in 15 countries. They reported moderate prior voice assistant experience (frequency: $M = 3.75$; comfort: $M = 4.16$; experience:

$M = 3.44$; 1–5 scales). Mean study completion time was 13.5 min ($SD = 4.3$, $Mdn = 13.5$). Table 2 summarizes the ASAQ results.

Table 2: ASAQ evaluation results ($n = 32$). Higher values indicate more favorable perception.

Construct	M	SD	95 % CI	α
Natural Behaviour	−0.21	1.77	[−0.82, 0.40]	.870
Agent's Coherence	1.85	1.26	[1.42, 2.29]	.879

NB was near-neutral ($M = -0.21$; CI spans zero), indicating that participants did not perceive the conversation as strongly natural or artificial. AC was positive ($M = 1.85$; CI entirely above zero), suggesting that responses were generally perceived as logical and consistent. Both constructs showed acceptable internal consistency ($\alpha > .87$). Relative to the ASAQ representative set 2024, the agent placed in the *moderate* range (50th–75th percentile) on both constructs, suggesting perceived quality within the range of existing agents despite being an early-stage open-source system. The coherence score is consistent with prior work identifying logical consistency as a key quality dimension for health conversational agents [9, 26]. Across all 422 turns in the 32 completed sessions, we report two latency metrics. Median time-to-first-audio (TTFA), from end of user speech to onset of the backchannel acknowledgment, was 1.5 s ($M = 1.6$ s, $P_{25} = 1.3$ s, $P_{75} = 1.7$ s, $P_5 = 1.2$ s, $P_{95} = 2.6$ s). Median time-to-substantive-audio (TTSA), from end of user speech to onset of the agent's substantive response, was 4.2 s. The interval between TTFA and TTSA had a median of 2.7 s; by architectural design, this interval covers backchannel playback together with the parallel execution of safety arbitration and substantive response generation. The tight TTFA interquartile range (0.4 s) indicates consistent backchannel responsiveness, while the P_{95} of 2.6 s suggests occasional spikes, likely attributable to model cold starts or GPU congestion during periods of concurrent usage. Interruption events, barge-in attempts, perceived delay, and exact bridge playback duration were not directly logged in this deployment and could not be quantified.

Qualitatively, participants noted accurate speech recognition and relevant exercise recommendations. The most frequent concern was premature turn-taking (approx. one third of participants), likely due to voice activity detection sensitivity. Additional suggestions included a more natural-sounding voice and more detailed advice. The near-zero naturalness score likely reflects the inherent challenge of voice-based interaction, where latency and interruptions are more salient than in text-based systems [8]. The qualitative turn-taking feedback supports this interpretation, pointing to voice activity detection as the primary factor limiting perceived naturalness.

5 Discussion and Future Work

Voice agents intended for clinical use must serve two audiences simultaneously. For users, the agent should feel like an autonomous, empathetic coach. For researchers, it must remain a transparent, reproducible system. STELLA offers an architectural path toward reconciling these requirements, and our deployment demonstrates its feasibility in a non-clinical setting. Stage 1 sustains conversational *presence* through backchanneling, achieving a median time-to-first-audio of 1.5 s. While this exceeds the ∼200 ms gap typical of human turn-taking [21], it falls safely within the *low*-latency condition where filler-based acknowledgments preserve conversational quality [27]. Furthermore, moderate delays can sometimes also increase perceived humanness [15]. A tight interquartile range (0.4 s) confirms this responsiveness is a stable property of the pipeline. Complementing this presence, Stage 4 ensures a consistent coaching *identity* while Stage 3 keeps action selection auditable by separating it from the generative model. By open-sourcing this architecture, STELLA decouples interaction design from safety infrastructure, empowering researchers to prototype coaching strategies while retaining inspectable control.

Our evaluation establishes system deployability and moderate conversational quality, but highlights several limitations. First, the study relies on a single use case (physical activity coaching) tested on a non-clinical Prolific sample, meaning it does not assess clinical appropriateness or therapeutic outcomes. Second, while feasibility is clear, we have not conducted ablation studies to empirically show whether this parallel architecture also outperforms simpler alternatives in perceived coherence or latency.

A closer review of participant dropouts reveals practical insights for future deployments. Of 46 consented participants, 14 did not complete the study, with 7 initially attributed to technical issues. Log inspections showed that 4 of these 7 sessions reached our 7-minute session cap before completion. Two of the three remaining sessions point to concrete learning opportunities for the platform: one in which the STT component returned an unusable transcript at a key turn, and one in which the participant's local audio hardware dropped out mid-conversation. For the third case, a survey response reported a technical error but no corresponding server log could be found on the platform, most likely reflecting an external participant-side issue. These two platform-side cases highlight concrete areas, namely STT robustness and resilience to client-side disruptions, where future iterations of the platform can be strengthened.

Qualitative findings point to the interaction layer as the true system bottleneck. Specifically, Voice Activity Detection (VAD) sensitivity and turn-taking calibration caused more conversational friction for users than the LLM-based reasoning. The current reliance on silence-based endpointing without barge-in support risks misinterpreting natural pauses. This is particularly problematic for populations with word-finding difficulties, such as individuals with early dementia, where retrieval pauses are frequent and lengthy.

A natural next step is a clinical deployment in Cognitive Stimulation Therapy for individuals with early dementia. Prior work has already established the feasibility of voice-based delivery [35, 36] and gathered crucial stakeholder perspectives [37] in this domain. Adapting STELLA for this population will require advancing turn-taking detection beyond pure silence-based endpointing [31]. Incorporating multimodal cues, such as visual attention signals, will be essential to properly distinguish intentional conversational yields from symptom-related delays.

Acknowledgments

This study was supported by the Swiss National Science Foundation (Grant No. 10002915) and the University Medicine Zurich (UMZH) technology initiative PRECIOUS, a Design and Trial Service for Precision Digital Therapeutics. Claude Opus 4.6 was used during the drafting process of Section 3 to enhance the clarity, readability, and structural flow. The authors meticulously reviewed all suggested edits and assume full responsibility for the final content.

Conflict of Interest

FM, RV, and TK are affiliated with the Centre for Digital Health Interventions (CDHI), a joint initiative of the Institute for Implementation Science in Health Care, University of Zurich; the Department of Management, Technology, and Economics, ETH Zurich; and the Institute of Technology Management and School of Medicine, University of St. Gallen. CDHI is funded in part by the Swiss health insurer CSS and the Austrian healthcare provider (and corporate start-up of UNIQA) Mavie Next, and Novo Nordisk Switzerland. TK is a cofounder of Pathmate Technologies, a university spin-off company that develops digital clinical pathways. TK has not held any shares in the company since 2023 and has never had an operational role within the company. However, neither CSS, Mavie Next (UNIQA), Novo Nordisk, nor Pathmate Technologies were involved in this study. The authors declare no other potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

[1] Mahyar Abbasian, Iman Azimi, Amir M Rahmani, and Ramesh Jain. 2025. Conversational health agents: a personalized large language model-powered agent framework. *JAMIA Open* 8, 4 (07 2025), ooaf067. arXiv:https://academic.oup.com/jamiaopen/article-pdf/8/4/ooaf067/63674695/ooaf067.pdf doi:10.1093/jamiaopen/ooaf067

[2] Lisa A Auster-Gussman, Kimberly G Lockwood, Sarah A Graham, Viveka Pitter, and OraLee H Branch. 2022. Engagement in Digital Health App-based prevention programs is associated with weight loss among adults age 65+. *Frontiers in digital health* 4 (2022), 886783. doi:10.3389/fdgth.2022.886783

[3] Timothy Bickmore, Amanda Gruber, and Rosalind Picard. 2005. Establishing the computer–patient working alliance in automated health behavior change interventions. *Patient Education and Counseling* 59, 1 (2005), 21–30. doi:10.1016/j.pec.2004.09.008

[4] Timothy W. Bickmore, Lisa Caruso, Kerri Clough-Gorr, and Tim Heeren. 2005. 'It's just like you talk to a friend' relational agents for older adults. *Interacting with Computers* 17, 6 (2005), 711–735. doi:10.1016/j.intcom.2005.09.002 HCI and the Older Population.

[5] Bland AI. 2024. Bland AI: Enterprise Voice Agent Infrastructure. https://www.bland.ai Commercial voice agent platform. Accessed March 2026.

[6] Harrison Chase. 2023. LangChain: Building Applications with LLMs through Composability. https://github.com/langchain-ai/langchain Accessed March 2026.

[7] Srinivas Chennabasappa et al. 2025. *LlamaFirewall: An Open Source Guardrail System for Building Secure AI Agents.* Technical Report. Meta AI Research.

[8] Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, and Benjamin R Cowan. 2019. The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers* 31, 4 (09 2019), 349–371. arXiv:https://academic.oup.com/iwc/article-pdf/31/4/349/33525046/iwz016.pdf doi:10.1093/iwc/iwz016

[9] Kerstin Denecke. 2023. Framework for Guiding the Development of High-Quality Conversational Agents in Healthcare. *Healthcare* 11, 8 (2023), 1061. doi:10.3390/healthcare11081061

[10] Qiuxue Du, Yongliang Ren, Ze-long Meng, Han He, and Shasha Meng. 2025. The Efficacy of Rule-Based Versus Large Language Model-Based Chatbots in Alleviating Symptoms of Depression and Anxiety: Systematic Review and Meta-Analysis. *J Med Internet Res* 27 (4 Dec 2025), e78186. doi:10.2196/78186

[11] ElevenLabs. 2024. ElevenLabs Conversational AI: Real-Time Voice Agent Platform. https://elevenlabs.io/conversational-ai Commercial platform, sub-100ms latency, 32+ languages. Accessed March 2026.

[12] European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L series. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[13] Siska Fitrianie, Merijn Bruijnes, Amal Abdulrahman, and Willem-Paul Brinkman. 2025. The Artificial Social Agent Questionnaire (ASAQ) — Development and evaluation of a validated instrument for capturing human interaction experiences with artificial social agents. *International Journal of Human-Computer Studies* 199 (2025), 103482. doi:10.1016/j.ijhcs.2025.103482

[14] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Ment Health* 4, 2 (06 Jun 2017), e19. doi:10.2196/mental.7785

[15] Ulrich Gnewuch, Stefan Morana, Marc Adam, and Alexander Maedche. 2018. Faster is Not Always Better: Understanding the Effect of Dynamic Response Delays in Human-Chatbot Interaction. In *Proceedings of the 26th European Conference on Information Systems (ECIS).* Association for Information Systems, Portsmouth, UK, 113. https://aisel.aisnet.org/ecis2018_rp/113

[16] Xinhui Hu, Smit Desai, Morgan Lundy, and Jessie Chin. 2025. Beyond Functionality: Co-Designing Voice User Interfaces for Older Adults' Well-being. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI '25).* Association for Computing Machinery, New York, NY, USA, Article 64, 15 pages. doi:10.1145/3719160.3736613

[17] Yining Hua, Steve Siddals, Zilin Ma, Isaac Galatzer-Levy, Winna Xia, Christine Hau, Hongbin Na, Matthew Flathers, Jake Linardon, Cyrus Ayubcha, and John Torous. 2025. Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: a systematic review. *World Psychiatry* 24, 3 (2025), 383–394. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/wps.21352 doi:10.1002/wps.21352

[18] Hume AI. 2024. Hume AI: Empathic Voice Interface. https://www.hume.ai Commercial voice AI platform. Accessed March 2026.

[19] Daniel Jaroslawicz, Brendan Whiting, Parth Shah, and Karime Maamari. 2025. How Many Instructions Can LLMs Follow at Once? arXiv:2507.11538 [cs.AI] https://arxiv.org/abs/2507.11538 Introduces the IFScale benchmark; reports that instruction-following accuracy degrades as instruction density increases, with smaller and non-reasoning models exhibiting the steepest declines.

[20] Matthew Jörke, Shardul Sapkota, Lyndsea Warkenthien, Niklas Vainio, Paul Schmiedmayer, Emma Brunskill, and James A. Landay. 2025. GPTCoach: Towards LLM-Based Physical Activity Coaching. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25).* Association for Computing Machinery, New York, NY, USA, Article 993, 46 pages. doi:10.1145/3706598.3713819

[21] Stephen C. Levinson and Francisco Torreira. 2015. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology* 6 (2015), 731. doi:10.3389/fpsyg.2015.00731

[22] Qiuxian Li and Haiou Wei. 2025. Factors influencing older adults' adoption of AI voice assistants: extending the UTAUT model. *Frontiers in Psychology* 16 (2025), 1618689. doi:10.3389/fpsyg.2025.1618689

[23] Guan-Ting Lin, Jiachen Lian, Tingle Li, Qirui Wang, Gopala Anumanchipalli, Alexander H. Liu, and Hung yi Lee. 2025. Full-Duplex-Bench: A Benchmark to Evaluate Full-duplex Spoken Dialogue Models on Turn-taking Capabilities. arXiv:2503.04721 [cs.CL] https://arxiv.org/abs/2503.04721

[24] LiveKit Inc. 2024. LiveKit Agents: Open Source Framework for Building Voice AI Agents. https://docs.livekit.io/agents/ Open-source agent framework built on LiveKit WebRTC. Accessed March 2026.

[25] LiveKit Inc. 2024. LiveKit: Open Source WebRTC Infrastructure for Real-Time Audio and Video. https://github.com/livekit/livekit Accessed March 2026.

[26] Laura Martinengo, Xiaowen Lin, Ahmad Ishqi Jabir, Tobias Kowatsch, Rifat Atun, Josip Car, and Lorainne Tudor Car. 2023. Conversational Agents in Health Care: Expert Interviews to Inform the Definition, Classification, and Conceptual Framework. *Journal of Medical Internet Research* 25 (2023), e50767. doi:10.2196/50767

[27] Mykola Maslych, Mohammadreza Katebi, Christopher Lee, Yahya Hmaiti, Amirpouya Ghasemaghaei, Christian Pumarada, Janneese Palmer, Esteban Segarra Martinez, Marco Emporio, Warren Snipes, Ryan P. McMahan, and Joseph J. LaViola Jr. 2025. Mitigating Response Delays in Free-Form Conversations with LLM-powered Intelligent Virtual Agents. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI '25).* Association for Computing Machinery, New York, NY, USA, Article 49, 15 pages. doi:10.1145/3719160.3736636

[28] Selina Meyer and David Elsweiler. 2025. LLM-based conversational agents for behaviour change support: A randomised controlled trial examining efficacy, safety, and the role of user behaviour. *International Journal of Human-Computer Studies* 200 (2025), 103514. doi:10.1016/j.ijhcs.2025.103514

[29] NVIDIA. 2024. NeMo Guardrails: An Open-Source Toolkit for Adding Programmable Guardrails to LLM-Based Conversational Systems. https://github.

com/NVIDIA-NeMo/Guardrails Accessed March 2026.

[30] Stefan Palan and Christian Schitter. 2018. Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17 (2018), 22–27. doi:10.1016/j.jbef.2017.12.004

[31] Gabriel Skantze. 2021. Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech & Language* 67 (2021), 101178. doi:10.1016/j.csl.2020.101178

[32] Guangzhi Sun, Xiao Zhan, and Jose Such. 2024. Building Better AI Agents: A Provocation on the Utilisation of Persona in LLM-based Conversational Agents. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces* (Luxembourg, Luxembourg) *(CUI '24)*. Association for Computing Machinery, New York, NY, USA, Article 35, 6 pages. doi:10.1145/3640794.3665887

[33] Tanh Quang Tran, Tobias Langlotz, Jacob Young, Thomas W. Schubert, and Holger Regenbrecht. 2024. Classifying Presence Scores: Insights and Analysis from Two Decades of the Igroup Presence Questionnaire (IPQ). *ACM Trans. Comput.-Hum. Interact.* 31, 5, Article 61 (Nov. 2024), 26 pages. doi:10.1145/3689046

[34] Vapi. 2024. Vapi: Voice AI for Developers. https://vapi.ai Commercial voice agent platform. Accessed March 2026.

[35] Rasita Vinay, Nora Camilla Tommila, Mathias Schlögl, Stefan Klöppel, Nikola Biller-Andorno, and Tobias Kowatsch. 2024. GRACE: Towards an Embodied Voice Assistant for Improving Quality of Life by Leveraging Elements of Cognitive Stimulation Therapy. In *Proceedings of Mensch Und Computer 2024* (Karlsruhe, Germany) *(MuC '24)*. Association for Computing Machinery, New York, NY, USA, 560–565. doi:10.1145/3670653.3677521

[36] Rasita Vinay, Ekaterina Uetova, Nora Camilla Tommila, Nikola Biller-Andorno, and Tobias Kowatsch. 2025. A Hybrid Rule- and Large Language Model-Based Embodied Voice Assistant (GRACE) for Cognitive Stimulation in Older Adults: Usability Study Assessing Technical Feasibility, Technology Acceptance, and Working Alliance. *JMIR Aging* 8 (2025), e76489. doi:10.2196/76489

[37] Florianne Walliser, Nikola Biller-Andorno, Tobias Kowatsch, and Rasita Vinay. 2025. Evaluating the GRACE Voice Assistant for Dementia Care Among Caregivers and Healthcare Professionals: An Interview Study. *medRxiv* 0, 0 (2025), 2025.10.08.25337586. doi:10.1101/2025.10.08.25337586 Preprint.

[38] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. arXiv:2308.08155 [cs.AI] https://arxiv.org/abs/2308.08155

[39] Ziqi Yang, Xuhai Xu, Bingsheng Yao, Ethan Rogers, Shao Zhang, Stephen Intille, Nawar Shara, Guodong Gordon Gao, and Dakuo Wang. 2024. Talk2Care: An LLM-based Voice Assistant for Communication between Healthcare Providers and Older Adults. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 73 (May 2024), 35 pages. doi:10.1145/3659625