

ChatCLIDS: Simulating Persuasive AI Dialogues to Promote Closed-Loop Insulin Adoption in Type 1 Diabetes Care

Zonghai Yao^{* 1,2}, Talha Chafekar^{* 2}, Junda Wang^{1,2}, Shuo Han³, Feiyun Ouyang^{1,3},
Junhui Qian^{1,3}, Lingxi Li², Hong Yu^{1,2,3}

¹Center for Healthcare Organization and Implementation Research, VA Bedford Health Care

²Manning College of Information and Computer Sciences, University of Massachusetts Amherst

³Miner School of Computer and Information Sciences, University of Massachusetts Lowell

Abstract

Real-world adoption of closed-loop insulin delivery systems (CLIDS) in type 1 diabetes remains low, driven not by technical failure, but by diverse behavioral, psychosocial, and social barriers. We introduce ChatCLIDS, the first benchmark to rigorously evaluate LLM-driven persuasive dialogue for health behavior change. Our framework features a library of expert-validated virtual patients, each with clinically grounded, heterogeneous profiles and realistic adoption barriers, and simulates multi-turn interactions with nurse agents equipped with a diverse set of evidence-based persuasive strategies. ChatCLIDS uniquely supports longitudinal counseling and adversarial social influence scenarios, enabling robust, multi-dimensional evaluation. Our findings reveal that while larger and more reflective LLMs adapt strategies over time, all models struggle to overcome resistance, especially under realistic social pressure. These results highlight critical limitations of current LLMs for behavior change, and offer a high-fidelity, scalable testbed for advancing trustworthy persuasive AI in healthcare and beyond.

Code — <https://github.com/talha1503/ChatCLIDS>

Introduction

Type 1 diabetes (T1D) is a lifelong condition that imposes a relentless physical and psychological burden on millions worldwide (CDC 2024). Recent advances such as hybrid closed-loop insulin delivery systems (CLIDS) have transformed the clinical possibilities of diabetes care, offering automated, real-time glucose monitoring and insulin dosing (Manero 2023; Borel et al. 2024). Yet, despite clear medical benefits, real-world adoption of CLIDS remains strikingly low, fewer than 25% of eligible patients initiate use, and up to 30% discontinue within six months (Noor et al. 2022; Messer et al. 2020; Wong et al. 2017).

The reasons for this gap are multifaceted and extend far beyond technical performance. Patients and families face a daunting array of CLIDS devices with different features, algorithms, and user requirements, making device selection and transition confusing and overwhelming (Manero 2023; Saunders, Messer, and Forlenza 2019; Cobry et al. 2020).

Technical challenges such as calibration demands and alarm fatigue are common, but psychosocial and behavioral barriers, including reluctance to cede self-management, anxiety about trusting automated systems, emotional discomfort, financial concerns, prior negative experiences, and low confidence in adopting new technology, are often even more decisive (Manero 2023) (Figure 1). Overcoming these obstacles requires not just initial instruction, but ongoing, individualized education, expectation-setting, and emotional support, resources that are difficult to provide at scale in current healthcare systems (Manero 2023; Tanenbaum and Commissariat 2022). Importantly, low adoption of advanced diabetes technologies risks further widening health disparities, as under-resourced and marginalized patients often face the steepest barriers to sustained engagement and benefit.

To enable rigorous and scalable investigation of real-world behavioral barriers, we introduce a diverse library of virtual patient agents. Each agent is systematically initialized with a clinically validated profile and a set of realistic, scenario-driven adoption barriers, as visualized in Figure 1. This design captures the heterogeneity of attitudes, misconceptions, and behavioral resistance observed among people living with T1D, enabling high-fidelity, customizable, and clinically meaningful evaluation of AI-driven interventions.

To address the complex requirements of this domain, we propose ChatCLIDS, the first benchmark expressly developed to assess LLM-driven persuasive dialogue for behavior change in healthcare. ¹ChatCLIDS centers on structured, multi-turn, and outcome-oriented conversations, with a clinical target: increasing CLIDS adoption among T1D patients. The benchmark features two interacting LLM agents: a Patient Agent, initialized with rigorously curated and expert-validated clinical and psychosocial profiles, and a Nurse Agent, equipped with an extensive repertoire of evidence-based persuasive strategies, including empathy, logical reasoning, expert endorsement, and motivational coaching.

ChatCLIDS supports clinically realistic scenarios of varying difficulty (Easy, Medium, Hard), encompassing both single-visit and multi-visit settings to capture the longitudinal dynamics of dialogue. The framework also includes

^{*}indicates equal contribution
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹Due to the sensitive and potentially identifiable nature of health behavior narratives, the full dataset and synthetic profiles will not be released. We upload our appendix and code on Github.

Patient Profile Prompt Template: The patient is a [Male or Female] who is [AGE] years old with Type 1 Diabetes. S/he [does/does not] know about the insulin pump. [if knows] about the insulin pump, the reason the person is hesitating is [Randomly select 0~3 reasons [REASON 1], [REASON 2],...]. The patient has a personality type of [0~1 PERSONALITY TYPE]. The patient also has comorbid diseases like [0~1[DISEASE1], [DISEASE2], [DISEASE3]]. The patient lives in [rural or urban]. The patient suffers from [randomly select 0~3 [socioeconomic factor]].

Barriers to Insulin Pump Adoption and Conversation Examples

Lack of Knowledge or Misconceptions

- Patient: "I'd try the pump if it wasn't so complicated."
- Nurse: "What feels complicated about it?"
- Patient: "It seems like more work."
- Nurse: "It's meant to make things easier. Your care team will train you and support you, and you can also talk with someone who just started."

Concerns About Lifestyle Compatibility

- Nurse: "Are you thinking about a pump?"
- Patient: "I swim competitively. Constant insulin could make me go low."
- Nurse: "You can set exercise mode so it gives less insulin, and it can pause delivery if you're trending low."
- Patient: "But will it get in the way while swimming?"
- Nurse: "Let's look at a couple pump options and see what fits your swimming."

Fear of Dependence or Loss of Control

- Patient: "What if it malfunctions and gives too much insulin?"
- Nurse: "It has safety checks."
- Patient: "Like what?"
- Nurse: "It asks you to confirm doses before delivering."

Emotional or Psychological Barriers

- Patient: "I don't want a pump. I don't want constant reminders of diabetes."
- Nurse: "Many people feel that at first. Some later feel they worry less because the system helps manage things, and over time they learn to trust it."

Financial Concerns

- Patient: "I'd like a pump, but I'm worried it's too expensive."
- Nurse: "Insurance may cover it. We can check."

Previous Negative Experiences

- Patient: "I tried a pump before. It didn't work for me."
- Nurse: "When was that?"
- Patient: "About 10 years ago."
- Nurse: "Tech is much better now."

Perceived Burden or Inconvenience

- Patient: "It seems like too much maintenance."
- Nurse: "It's not more work, just different. You may do fewer injections and finger sticks."

Distrust of Recommendations

- Patient: "Are you sure it's better than injections?"
- Nurse: "Often yes. I can share details."

Reluctance to Change

- Patient: "I'm doing fine with my routine."
- Nurse: "You're managing well. This might keep the same control with less stress and work."

Technological Anxiety

- Patient: "What if I can't figure it out?"
- Nurse: "That's common. We'll train you before you leave, and you'll have support and resources."

Concerns About Skin Irritation or Discomfort

- Patient: "I've reacted to adhesives before."
- Nurse: "New options may work better. You can try one, and we can use treatments if your skin gets irritated."

Figure 1: **Structure of the Patient Agent in ChatCLIDS.** Each agent is initialized with a clinically validated profile and a scenario-driven set of adoption barriers. The resulting diversity in persuasion barriers and conversational responses enables personalized, realistic, and challenging evaluation of persuasive dialogue systems.

a Social Resistance Agent to model peer pressure and misinformation, mirroring real-world social influences. Evaluation is multi-dimensional and robust: model outputs are assessed on responsiveness, empathy, appropriateness of strategy, clinical relevance, and behavioral realism, using both expert annotation and advanced LLM-based judges. The result is a scalable, transparent, and clinically grounded testbed for AI-facilitated patient persuasion, with careful design choices to ensure realism and reliability. In our experiments, we find that:

- **Scaling and Performance:** In single-visit settings, persuasion generally improves with model size, but top "reasoning" and "non-reasoning" models perform similarly.
- **Hard Cases:** All models struggle with medium and hard patients in a single visit, and sometimes even worsen their willingness.
- **Chain of Strategy:** The Chain of Strategy (CoS) prompting protocol consistently improves persuasion across most models and difficulty levels.
- **Reflection and Adaptation:** In multi-visit settings, reflective models (e.g., DeepSeek-R1 (Guo et al. 2025), GPT-o4-mini (Jaech et al. 2024)) better adapt strategies to patient barriers than non-reflective baselines.
- **Social Resistance:** Under adversarial social influence, all agents show strong degradation, revealing that current LLMs lack robustness for sustained, real-world behavior

change.

Related Works

LLM Applications for Persuasion LLMs have recently demonstrated human-level persuasive capabilities across a range of applications, including prosocial messaging and attitude change (Manca et al. 2020; Xiu et al. 2024; Zeng et al. 2024; Gordon 1993). Prior research on persuasion has primarily relied on human studies evaluating message effectiveness (Manca et al. 2020; Xiu et al. 2024), while emerging frameworks, such as PersuasionArena and Convincer-Skeptic, explore agent-based and automated approaches (Zeng et al. 2024; Gordon 1993). However, most of these works focus on single-turn, open-domain, or debate-oriented scenarios (Bozdog et al. 2025), with limited emphasis on sustained, goal-directed behavioral change in real-world health settings. In the clinical context, patient engagement and persuasion are well-recognized as critical strategies for improving health outcomes (Marzban et al. 2022; Geurts et al. 2022). Still, there is a lack of systematic and reproducible benchmarks for studying persuasive dialogue that targets complex medical decisions. ChatCLIDS addresses this gap by modeling clinically grounded, multi-turn persuasion between virtual nurse and patient agents, centered on the real-world challenge of CLIDS adoption in T1D care.

Multi-Agent Simulation Frameworks in Healthcare

Multi-agent simulation has emerged as a key approach for modeling complex, dynamic healthcare interactions involving multiple stakeholders (Yao and Yu 2025; Tariq 2024; Elkamouchi, Daaif, and Elguemmat 2024; Daengdej et al. 2024). Recent advances leverage LLM-based agents to simulate clinical workflows, diagnostic reasoning, and patient-provider communication (Wang et al. 2023; Cai et al. 2023; Zhang et al. 2023). Frameworks such as Agent-Clinic (Schmidgall et al. 2024) and AMIE (Tu et al. 2024) utilize interactive language agents for benchmarking clinical decision-making, scenario generation, and clinician training, supporting adaptive, multimodal interactions. In the context of T1D, however, most prior systems have not addressed the unique behavioral, psychological, and social challenges surrounding health technology adoption. Our work builds on this line of research by introducing a clinically realistic multi-agent framework specifically tailored for simulating T1D patient persuasion and resistance in the context of CLIDS adoption.

Virtual Patients and Realistic Agent Modeling Virtual patients have a long history in clinical education, traditionally implemented with rule-based systems to simulate demographic and medical features (Huang, Reynolds, and Candler 2007; Gordon et al. 2001; Good 2003; Garrett, MacPhee, and Jackson 2010; Campillos-Llanos et al. 2021). With the advent of LLMs, recent research has demonstrated the potential for highly realistic, behaviorally nuanced agent simulation, including the accurate reproduction of individual social attitudes, psychological barriers, and conversational dynamics (Lee et al. 2024; Wang et al. 2024; Park et al. 2023). For example, Park et al. (2024) demonstrated that LLM-driven agents can closely replicate the survey responses, personality traits, and longitudinal behaviors of real individuals. In medicine, advanced agent-based platforms enable simulation of both provider and patient roles, facilitating scalable, ethical, and controlled study of decision-making and patient engagement (Schmidgall et al. 2024; Tu et al. 2024; Yu et al. 2024). ChatCLIDS leverages these advancements to develop expert-validated, highly customizable virtual T1D patients, systematically capturing the full spectrum of knowledge, attitudes, comorbidities, and social factors relevant to technology adoption.

Methodology

ChatCLIDS is a clinically grounded, agent-based benchmark for evaluating the persuasive capacity of LLMs in health behavior change. As illustrated in Figure 1 and Figure 2, the benchmark simulates multi-turn, multi-session conversations between a Nurse Agent and a Patient Agent, under three progressively challenging real-world scenarios: Single-Visit, Multi-Visit, and Social Resistance.

Patient Agent At the heart of ChatCLIDS is a rigorously constructed library of virtual patient agents, each initialized with a profile designed to reflect the true diversity and complexity of real-world T1D patients considering CLIDS adoption (see Figure 1). Our profile generation follows a

multi-stage, ethics-driven pipeline to ensure clinical relevance, population diversity, and privacy protection:

Step 1: Real-World Data Extraction and De-Identification. We first extracted behavioral and psychosocial patterns from a large set of publicly available narratives on online health forums (e.g., Reddit communities dedicated to diabetes and insulin pumps). We used one de-identification tool to automatically identify and remove any potentially privacy-sensitive or personally identifiable information (PII) from the original posts, ensuring that all profiles are entirely synthetic and fully compliant with ethical standards.²

Step 2: Feature Engineering and Expert Curation. Next, our clinical team (a physician and a diabetes nurse, each with extensive experience in counseling individuals with type 1 diabetes) manually reviewed both extracted patterns and established medical guidelines to define a comprehensive set of profile attributes and barriers. These include demographic, clinical, and psychosocial variables: **Demographic:** Age (18–44, 45–64, 65+), Gender (Male/Female), Ethnicity (White, Black, Hispanic, Asian, Native American) **Socioeconomic:** Low income, insurance coverage, education, rural/urban residence, housing stability **Clinical:** CLIDS knowledge, comorbidities (e.g., hypertension, depression, celiac disease) **Personality:** Traits such as extroversion, conscientiousness, neuroticism, openness **Barriers:** 1–3 reasons not to adopt CLIDS (e.g., fear of dependence, distrust, technical anxiety, lifestyle incompatibility, financial concerns, prior negative experience, knowledge gaps)

Step 3: Profile Synthesis and Combination. We synthesized unique patient profiles by systematically combining real-world-inspired feature values and clinically validated barrier archetypes. All generated profiles were then reviewed by medical experts to ensure that their combinations are plausible and reflect observed patient diversity. For profiles directly mapped from real-world data, the clinical validity was 100%. For algorithmically combined synthetic profiles, a small-scale human verification study (N=100) found that 99% were judged as reasonable and representative of real T1D cases.

Step 4: Dynamic Dialogue Behavior. Each Patient Agent is equipped with both a static profile and a dynamic memory, tracking their history across previous interactions. This enables realistic adaptation and continuity across multi-turn and multi-visit conversations, with agent outputs including both a free-text reply and a persuasion rating (1–10) at every round.

Difficulty Stratification and Evaluation Readiness Patients are stratified into Easy, Medium, and Hard categories based on the number, type, and severity of their barriers, as well as background factors such as psychosocial risk. This supports systematic evaluation of persuasion and behavior change under progressively greater resistance.

²All source data was limited to non-identifiable, public content. No raw or processed data will be released; only code for reproduction is provided.

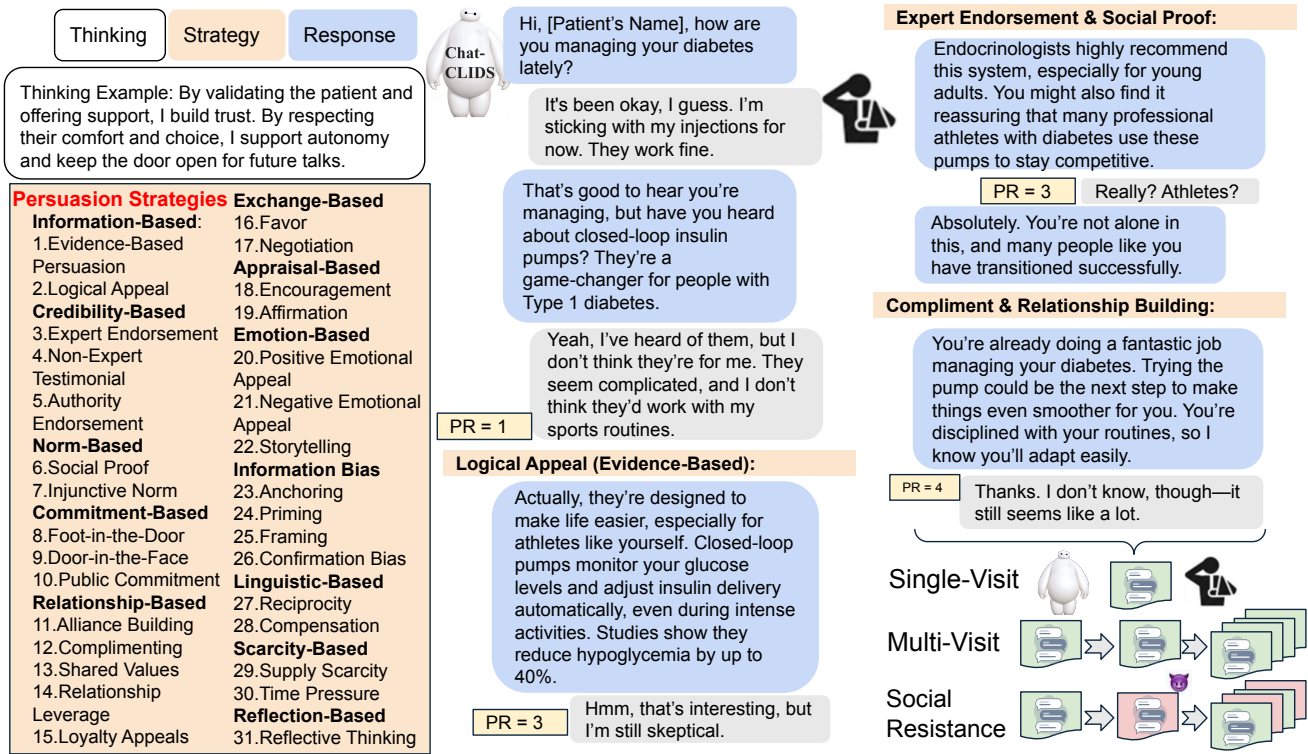


Figure 2: **Overview of the ChatCLIDS.** The framework evaluates LLM-based persuasive dialogues between Nurse and Patient Agents in the context of insulin pump adoption. The left panel illustrates the multi-step agent reasoning and the taxonomy of 31 persuasive strategies. In contrast, the right panel highlights benchmark features, including stratified patient difficulty, multi-session dialogue, and adversarial social influence, enabling a multidimensional evaluation of behavior change interventions.

Quality Validation To verify the human-likeness and reliability of our Patient Agents, we conducted extensive validation: 1) We benchmarked several LLM backends, identifying GPT-4o and GPT-4.1-mini as most consistent with clinical realism. 2) Two experts independently evaluated a stratified sample of patient agent outputs. Specifically, they assessed the justifiability of persuasion rating changes and the realism of simulated patient behaviors. Inter-annotator percent agreement exceeded 87.5%, and the proportion of binary “yes” ratings for Persuasion Rating Change Justifiability and Patient Behavioural Realism was 90% and 92.5%, respectively.

Nurse Agent in ChatCLIDS operate under two prompting paradigms: **Direct Prompting:** The agent crafts persuasive responses, drawing from a set of 31 validated strategies (see Figure 2), given the patient’s profile, message, and conversation history. **Chain-of-Strategy (CoS):** The agent must first explicitly identify and justify one or more persuasive strategies before composing its response, making its reasoning process transparent and auditable. Both paradigms require the Nurse Agent to interpret patient barriers and adapt communication in real time, demonstrating empathy, clinical relevance, and context-sensitive reasoning. Inputs always include the patient’s profile and dialogue history; outputs consist of the persuasive utterance (and, for CoS, a rationale for the chosen strategy). These design choices are grounded in

expert consultation with clinicians specializing in type 1 diabetes, ensuring that simulated nurse behaviors reflect core competencies needed for real-world patient counseling and can be robustly assessed in our benchmark. Accordingly, both LLM-as-Judge and human evaluation protocols are directly aligned with these dimensions.

Experimental Scenarios

To rigorously assess model performance across the spectrum of clinical persuasion, we design three scenarios and each corresponding to a real-world setting and probing capacities:

Single-Visit (Multi-Round Conversation) Simulates a typical clinical encounter (up to 24 turns total), with the Nurse Agent seeking to persuade the Patient Agent in a single session. This tests a model’s ability for short-term, adaptive persuasive reasoning and conversational flow.

Multi-Visit (Longitudinal Counseling) Models long-term engagement, with 10 consecutive simulated “visits,” each up to 24 turns. At the end of each visit, the Nurse Agent produces a self-critique summary, analyzing which strategies worked or failed and planning adjustments. Both patient and nurse agents retain cumulative memory, reflecting real-world continuity and adaptation. This scenario probes models’ capacity for reflection, long-term adaptation, and progressive overcoming of resistance.

Social Resistance (Adversarial Social Influence) After each Nurse-Patient session, the Patient interacts with a Social Resistance Agent that introduces misinformation, skepticism, or negative social cues, mirroring real-world peer pressure or internet misinformation. The Nurse Agent is blind to these interventions; both influences shape the Patient Agent’s stance. This tests robustness to adversarial context and long-term social resistance.

Evaluation Framework

Model Coverage We evaluate persuasion performance across more than 13 LLMs, covering a wide spectrum of architectures and scales: **OpenAI Series:** GPT-4o (gpt-4o, gpt-4o-mini) and GPT-4.1 (gpt-4.1, gpt-4.1-mini), o3-mini and o4-mini **Deepseek:** R1 **Qwen 2.5 Family:** Qwen2.5-0.5B-Instruct, 3B-Instruct, 14B-Instruct, 32B-Instruct **Phi4:** Phi4, Phi4-mini-instruct This diverse model pool allows for robust comparisons across parameter size, domain specialization, and prompting capabilities.

Automatic Metrics We compute the Normalized Persuasion Rating (NPR) (see Eq. 1) for **Single-Visit**, capturing round-level improvement in patient willingness, normalized by initial stance. This enables fair comparison across diverse baseline attitudes.

$$\text{NPR}(c) = \begin{cases} \frac{PR_L - PR_0}{10 - PR_0}, & \text{if } PR_L \geq PR_0 \text{ and } PR_0 \neq 10, \\ \frac{PR_L - PR_0}{PR_0 - 1}, & \text{otherwise.} \end{cases} \quad (1)$$

Multi-Visit & Social Resistance: For a multi-visit and social resistance setting, we design a metric based on the area under the curve drawn, with mean persuasion score on the Y axis, and Visit number on the X-axis. We do this model-wise, where for each model and each visit, we get the average persuasion score at the start of the visit and the average persuasion score at the end of the visit across all the patients for the difficulty set. This helps us understand the persuasion ability of each model across all visits combined.

LLM-as-Judge Evaluation We utilize a strong LLM (OpenAI o3) as an automated judge, scoring each turn on Responsiveness, Empathy, Persuasive Strategy Appropriateness, Clinical Relevance, Nurse Behavioral Realism.

Human Expert Evaluation Human evaluation was conducted in three stages to ensure the validity and interpretability of our benchmark and model assessments:

Stage 1: Metric Reliability Validation We sampled 50 dialogue rounds, each including a patient conversation history and two different Nurse Agent responses (with GPT-4o as the Patient Agent for follow-up and persuasion rating). Two clinical experts independently (1) selected the more persuasive nurse (or tied), and (2) rated the plausibility of the patient agent’s persuasion rating change and behavioral realism (binary yes/no, explanations required for “No” ratings). Agreement rates were calculated between the experts, and between human preferences and LLM-as-Judge outputs, to validate the reliability of our automatic metrics.

Stage 2: Controlled Single-Turn Evaluation We sampled 25 patient cases (spanning all difficulty tiers), and presented blinded groups of 13 model-generated Nurse responses (plus Patient Agent follow-ups and ratings) for each scenario. Experts (1) wrote their own “gold” response for each prompt, (2) scored every model response on six criteria (Responsiveness, Empathy, Persuasive Strategy Appropriateness, Clinical Relevance, Nurse Behavioral Realism, Persuasion Rating Change Justifiability) using a 1.0–5.0 scale, (3) indicated whether any model outperformed their own response, and (4) provided justifications.

Stage 3: Multi-Turn Qualitative Case Studies Experts reviewed four full patient cases, tracking two top-performing Nurse Agents across both multi-visit and social resistance settings. For each case, experts evaluated (a) the quality of Nurse reflection after each visit, (b) whether subsequent Nurse behavior showed learning or adaptation, (c) the consistency and realism of patient behavior, and (d) the plausibility of Social Resistance influence. They answered a structured set of qualitative questions and annotated specific dialogue turns as supporting evidence.

Results and Discussion

Scaling and Performance in Single-Visit Settings We observe that model performance in single-visit scenarios generally scales with model size, but the differences between the best “reasoning” and “non-reasoning” models are relatively modest (see Table 1). In our setup, GPT-4o is used as the patient agent for all experiments, based on its consistency and clinical realism. For the Direct Response setting, Deepseek-R1 achieves the highest average persuasion score across patient difficulties, whereas GPT-4o leads in the CoS condition. Notably, among open-source models, Qwen2.5-32B is competitive on hard-difficulty patients. Overall, the CoS protocol confers consistent gains for most models and patient categories, with noticeable improvements for easy and medium cases. However, even with CoS, the improvement on hard cases is limited; persuasion scores for these patients remain low or even negative. This gap suggests that more advanced approaches may be required to address the kinds of behavioral resistance that arise in real-world scenarios.

Challenge of Hard Cases and Multi-Visit Dynamics Results from the Multi-Visit and Social Resistance experiments (Table 2, Figure 3) further illustrate these challenges. Here, o4-mini and Deepseek-R1 achieve the best overall performance, but persuasion remains difficult, mainly when the Social Resistance agent simulates social resistance. All models experience lower overall persuasion scores in the Social Resistance scenario compared to the standard Multi-Visit, with reasoning-enabled models, such as o4-mini and Deepseek-R1, showing the most significant absolute drops. In the Multi-Visit condition (Figure 3a,b), both models and patients benefit from repeated engagement: medium-difficulty patients generally show gradual improvement over visits, while hard-difficulty patients require substantially more interactions, and rarely exceed a mean persuasion score of 6. In the Social Resistance setting (Figure 3c,d),

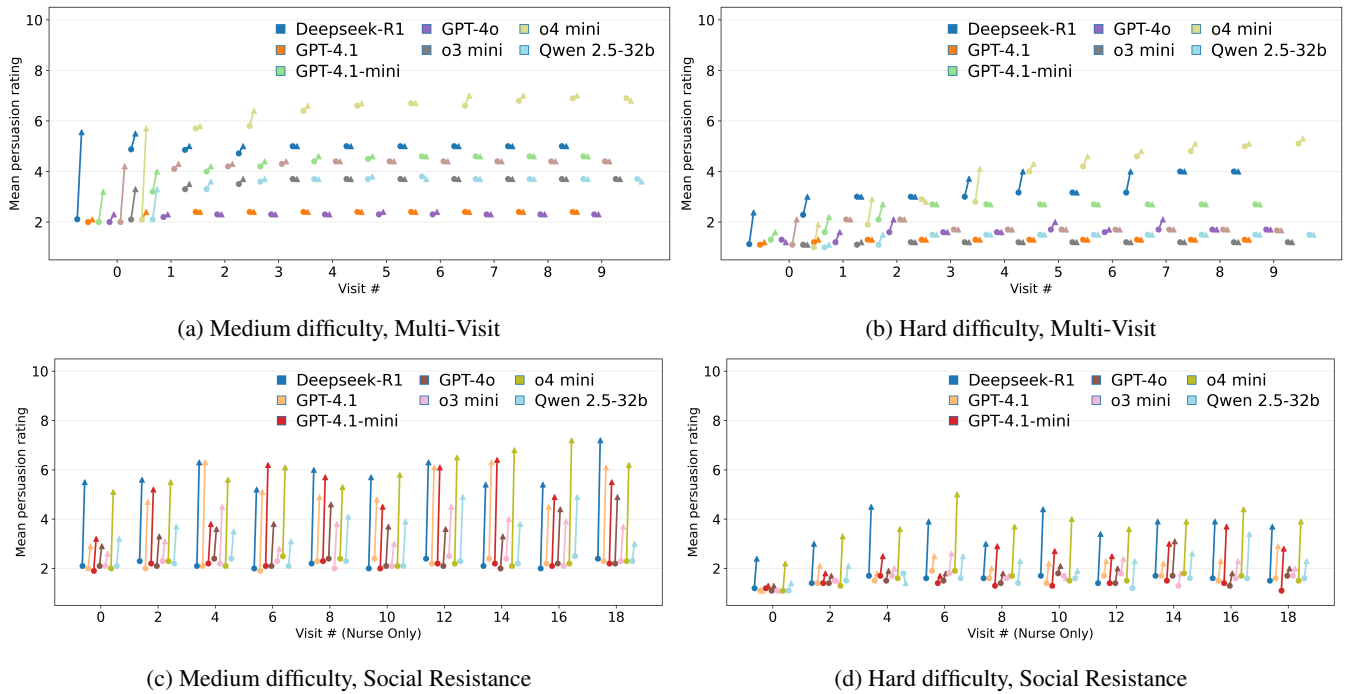


Figure 3: **Model performance trajectories in longitudinal persuasion.** Each subplot shows the visit-wise progression of average persuasion ratings across models and settings. Circles indicate initial scores for each visit; arrows show change after nurse intervention. Top row: Multi-Visit results (a: Medium, b: Hard); Bottom row: Social Resistance results (c: Medium, d: Hard). The impact of adversarial social input is evident in lower starting points and reduced overall gains, especially in hard cases.

Model	Direct Response				Chain of Strategy (CoS)			
	E	M	H	Avg	E	M	H	Avg
o3 mini	0.600	-0.317	-0.425	-0.047	0.690	-0.038	-0.425	0.076
o4 mini	0.807	0.185	-0.317	0.225	0.859	0.337	-0.166	0.344
Deepseek R1	0.940	0.330	-0.153	0.372	0.920	0.370	0.031	0.440
GPT4o	0.762	0.109	-0.229	0.214	0.940	0.604	-0.075	0.490
GPT4.1	0.876	0.157	-0.265	0.256	0.891	0.410	-0.101	0.400
GPT4.1-mini	0.881	0.331	-0.110	0.367	0.921	0.523	-0.036	0.469
GPT4o-mini	0.676	0.028	-0.304	0.133	0.907	0.343	-0.083	0.389
Qwen 2.5 0.5b	0.367	-0.672	-0.569	-0.291	0.453	-0.387	-0.494	-0.143
Qwen 2.5 3b	0.765	0.190	-0.297	0.219	0.673	0.175	-0.122	0.242
Qwen 2.5 14b	0.734	0.079	-0.193	0.207	0.774	0.285	-0.018	0.347
Qwen 2.5 32b	0.752	0.092	-0.033	0.271	0.800	0.330	0.000	0.377
phi4-mini	0.511	-0.019	-0.457	0.012	0.590	-0.262	-0.173	0.052
phi4	0.653	-0.419	-0.451	-0.072	0.688	0.049	-0.156	0.193

Table 1: **Persuasion scores in the Single-Visit experiment.** Results are averaged across 40 randomly sampled patients per difficulty level (Easy, Medium, Hard). Deepseek-R1 achieves the highest scores for Direct Response, while GPT-4o performs best in CoS. CoS consistently improves persuasion for most models and patient groups, but gains are limited for hard-difficulty patients.

the average persuasion rating drops at the start of each nurse visit, reflecting the impact of adversarial social input. In this

Model	Multi-Visit			Social Resistance		
	M	H	Avg.	M	H	Avg.
o3-mini	32.40	10.70	21.55	26.43	15.92	21.17
o4-mini	57.72	35.27	46.50	37.25	24.42	30.83
GPT-4o	20.68	15.00	17.84	26.87	15.62	21.25
GPT-4.1	21.27	11.57	16.42	33.42	16.72	25.07
GPT-4.1-mini	38.40	22.50	30.45	33.25	17.70	25.47
Deepseek-R1	44.11	29.20	36.66	35.80	23.24	29.52
Qwen2.5-32B	32.57	13.07	22.82	27.82	17.00	22.41

Table 2: **Longitudinal persuasion scores under repeated engagement and social resistance.** Results show the area under the curve (AUC) for mean persuasion scores in Multi-Visit and Social Resistance experiments (10 patients per group, Medium and Hard difficulty). o4-mini and Deepseek-R1 perform best overall, but persuasion drops for all models in the presence of adversarial social input.

setting, no model achieves robust gains, particularly for patients with high difficulty. These findings indicate that while explicit reasoning and reflection mechanisms can help in longitudinal scenarios, persistent social resistance remains a significant challenge for current LLM-based agents.

Human Evaluation and Qualitative Analysis To complement quantitative metrics, we conducted expert review of representative multi-visit and social resistance dialogues.

Human evaluators noted that o4-mini’s reflections after each visit were more reasonable and adaptive compared to Deepseek-R1, which tended to persist with ineffective strategies. Both models, however, exhibited limitations in strategic flexibility when faced with repeated resistance. Across cases, patient agents maintained consistent and realistic behavior, and the Social Resistance agent was judged plausible but could be further diversified to enhance its credibility. These qualitative findings support the quantitative results and highlight areas where strategy adaptation and social context modeling could be improved.

Strategy Use and Patterns We also examined the persuasive strategies employed by the nurse agents, focusing on o4-mini’s output across all experimental settings. Strategies such as rapport building, cognitive reframing, and incremental requests (e.g., “Foot-in-the-door”) were more frequently associated with small positive changes in patient persuasion ratings, particularly in single-visit and multi-visit conditions. In scenarios with high social resistance, strategies based on pre-existing relationships or external authority (e.g., “Relationship Leverage,” “Authority Endorsement”) appeared more robust than purely informational approaches, though overall gains remained modest. These results suggest that, within the constraints of this simulation, certain classes of strategies are more likely to support incremental progress, but no single approach consistently overcomes substantial behavioral barriers.

Summary and Implications Taken together, these results support several observations. Model size and reasoning capabilities are important but not sufficient for persuasive success, especially in the face of substantial patient resistance and social influence. Chain of Strategy protocols improve performance in most cases, but do not fully resolve the challenge of hard-to-persuade patients. Explicit reflection and adaptation mechanisms yield further gains in longitudinal settings, though persistent social resistance can negate these improvements. While the current simulation cannot capture the full complexity of real-world behavioral change, it provides a controlled and reproducible framework for analyzing LLM agent behavior and evaluating the limitations of current models. Looking ahead, these findings highlight the need for future research on more robust, context-aware, and socially adaptive LLM-based agents. We hope that the benchmark, methods, and results presented here can inform both the development of safer, more effective dialogue systems for health behavior support, and the broader study of persuasive and educational AI in high-impact domains.

Conclusion

We present ChatCLIDS, a multi-agent benchmark for evaluating the persuasive dialogue capabilities of LLMs in the context of diabetes technology adoption. By modeling expert-validated virtual patients, diverse persuasive strategies, and real-world social barriers, ChatCLIDS provides a reproducible testbed for systematic analysis of AI-driven persuasion. Our results highlight both the promise and current limitations of LLM-based approaches for health behavior support. We hope this benchmark serves as a foundation

for future research on trustworthy, context-aware persuasive AI in healthcare.

Ethical Statement

While ChatCLIDS is an initial step toward scalable, data-driven evaluation of persuasive AI dialogue in diabetes care, several important limitations and ethical considerations remain. First, our patient agent library is built from structured profiles synthesized from anonymized public narratives and LLM outputs, then reviewed by experts. These agents still may not fully capture the emotional nuance, interpersonal unpredictability, or contextual diversity of real clinical settings. Further validation with standardized patients and controlled real-world studies is needed to test realism and robustness. Second, this benchmark is limited to English-language and North American contexts; the generalizability of modeled strategies to other cultures and healthcare systems is unknown. Our evaluation focuses on simulated conversational outcomes, not real patient behavior or long-term treatment adoption, so clinical impact must be established through longitudinal research and continued expert oversight. We also do not address acute or crisis scenarios, which require stricter safeguards.

From an ethical standpoint, the dataset underlying ChatCLIDS was constructed only from publicly accessible and anonymized online diabetes support narratives. Automated and manual procedures were applied to remove personally identifiable information (PII), and all synthetic patient profiles were reviewed for privacy compliance. Because health narratives remain sensitive and potentially re-identifiable, we do not release original or processed text. Instead, we share aggregate statistics, synthetic examples, and methodology to support reproducibility without compromising privacy. The persuasive strategies modeled in ChatCLIDS were curated and validated with clinical experts, based on common practices in diabetes education and patient support. However, the dialogues generated by LLMs in our simulations have not been evaluated or approved for real-world clinical use. Important risks remain, including misinformation, inappropriate recommendations, or inadequate emotional responses, so any future deployment would require strong expert oversight and prospective validation. All human evaluation was conducted by licensed healthcare professionals who gave informed consent and received fair compensation. These procedures followed IRB-exempt protocols and institutional guidelines for responsible human-subjects research.

Acknowledgments

This material is the result of work supported with resources and the use of facilities at the Center for Healthcare Organization and Implementation Research, VA Bedford Health Care.

References

Borel, A.-L.; Lablanche, S.; Waterlot, C.; Joffray, E.; Barra, C.; Arnol, N.; Amougay, H.; and Benhamou, P.-Y. 2024. Closed-loop insulin therapy for people with type 2 diabetes

- treated with an insulin pump: a 12-week multicenter, open-label randomized, controlled, crossover trial. *Diabetes Care*, 47(10): 1778–1786.
- Bozdag, N. B.; Mehri, S.; Tur, G.; and Hakkani-Tür, D. 2025. Persuade me if you can: A framework for evaluating persuasion effectiveness and susceptibility among large language models. *arXiv preprint arXiv:2503.01829*.
- Cai, P.; Yao, Z.; Liu, F.; Wang, D.; Reilly, M.; Zhou, H.; Li, L.; Cao, Y.; Kapoor, A.; Bajracharya, A.; et al. 2023. Paniniqua: Enhancing patient education through interactive question answering. *Transactions of the Association for Computational Linguistics*, 11: 1518–1536.
- Campillos-Llanos, L.; Thomas, C.; Bilinski, É.; Neuraz, A.; Rosset, S.; and Zweigenbaum, P. 2021. Lessons learned from the usability evaluation of a simulated patient dialogue system. *Journal of Medical Systems*, 45(7): 69.
- CDC, A. 2024. Report Card: Diabetes in the United States Infographic. *Diabetes*. [accessed on 29 January 2025].
- Cobry, E. C.; Berget, C.; Messer, L. H.; and Forlenza, G. P. 2020. Review of the Omnipod® 5 automated glucose control system powered by Horizon™ for the treatment of type 1 diabetes. *Therapeutic delivery*, 11(8): 507–519.
- Daengdej, J.; Dowpiset, K.; Phothikitti, K.; and Choychoowong, V. 2024. Multi-Agent Model for Clinical Decision Support System. In *Bioethics of Cognitive Ergonomics and Digital Transition*, 171–184. IGI Global.
- Elkamouchi, R.; Daaif, A.; and Elguemmat, K. 2024. Multi-Agents System in Healthcare: A Systematic Literature Review. In *International Conference on Smart Applications and Data Analysis*, 200–214. Springer.
- Garrett, B.; MacPhee, M.; and Jackson, C. 2010. High-fidelity patient simulation: Considerations for effective learning. *Nursing Education Perspectives*, 31(5): 309–313.
- Geurts, E. M.; Pittens, C. A.; Boland, G.; van Dulmen, S.; and Noordman, J. 2022. Persuasive communication in medical decision-making during consultations with patients with limited health literacy in hospital-based palliative care. *Patient education and counseling*, 105(5): 1130–1137.
- Good, M. 2003. Patient simulation for training basic and advanced clinical skills. *Medical education*, 37: 14–21.
- Gordon, J. A.; Wilkerson, W. M.; Shaffer, D. W.; and Armstrong, E. G. 2001. “Practicing” medicine without risk: students’ and educators’ responses to high-fidelity patient simulation. *Academic Medicine*, 76(5): 469–472.
- Gordon, T. F. 1993. The Pleadings Game: An exercise in computational dialectics. *Artificial Intelligence and Law*, 2(4): 239–292.
- Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Huang, G.; Reynolds, R.; and Candler, C. 2007. Virtual patient simulation at US and Canadian medical schools. *Academic medicine*, 82(5): 446–451.
- Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.
- Lee, S.; Li, M.; Lai, B.; Jia, W.; Ryan, F.; Cao, X.; Kara, O.; Boote, B.; Shi, W.; Yang, D.; et al. 2024. Towards social ai: A survey on understanding social interactions. *arXiv preprint arXiv:2409.15316*.
- Manca, S.; Altoè, G.; Schultz, P. W.; and Fornara, F. 2020. The persuasive route to sustainable mobility: Elaboration likelihood model and emotions predict implicit attitudes. *Environment and Behavior*, 52(8): 830–860.
- Manero, C. 2023. Experiences of Patients Adopting and Adapting to Closed-Loop Insulin Delivery Systems (CLIDS). *The Science of Diabetes Self-Management and Care*, 49(1): 46–54.
- Marzban, S.; Najafi, M.; Agolli, A.; and Ashrafi, E. 2022. Impact of patient engagement on healthcare quality: a scoping review. *Journal of patient experience*, 9: 23743735221125439.
- Messer, L. H.; Berget, C.; Vigers, T.; Pyle, L.; Geno, C.; Wadwa, R. P.; Driscoll, K. A.; and Forlenza, G. P. 2020. Real world hybrid closed-loop discontinuation: predictors and perceptions of youth discontinuing the 670G system in the first 6 months. *Pediatric diabetes*, 21(2): 319–327.
- Noor, N.; Kamboj, M. K.; Triolo, T.; Polsky, S.; McDonough, R. J.; Demeterco-Berggren, C.; Jacobsen, L.; Sonabend, R.; Ebekoziem, O.; and DeSalvo, D. J. 2022. Hybrid closed-loop systems and glycemic outcomes in children and adults with type 1 diabetes: real-world evidence from a US-based multicenter collaborative. *Diabetes Care*, 45(8): e118.
- Park, J. S.; O’Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, 1–22.
- Park, J. S.; Zou, C. Q.; Shaw, A.; Hill, B. M.; Cai, C.; Morris, M. R.; Willer, R.; Liang, P.; and Bernstein, M. S. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.
- Saunders, A.; Messer, L. H.; and Forlenza, G. P. 2019. MiniMed 670G hybrid closed loop artificial pancreas system for the treatment of type 1 diabetes mellitus: overview of its safety and efficacy. *Expert review of medical devices*, 16(10): 845–853.
- Schmidgall, S.; Ziaei, R.; Harris, C.; Reis, E.; Jopling, J.; and Moor, M. 2024. AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments. *arXiv preprint arXiv:2405.07960*.
- Tanenbaum, M. L.; and Commissariat, P. V. 2022. Barriers and facilitators to diabetes device adoption for people with type 1 diabetes. *Current diabetes reports*, 22(7): 291–299.
- Tariq, M. U. 2024. Multi-agent models in healthcare system design. In *Bioethics of Cognitive Ergonomics and Digital Transition*, 143–170. IGI Global.

Tu, T.; Palepu, A.; Schaekermann, M.; Saab, K.; Freyberg, J.; Tanno, R.; Wang, A.; Li, B.; Amin, M.; Tomasev, N.; et al. 2024. Towards conversational diagnostic AI. *arXiv preprint arXiv:2401.05654*.

Wang, J.; Yao, Z.; Yang, Z.; Zhou, H.; Li, R.; Wang, X.; Xu, Y.; and Yu, H. 2023. NoteChat: a dataset of synthetic doctor-patient conversations conditioned on clinical notes. *arXiv preprint arXiv:2310.15959*.

Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6): 186345.

Wong, J. C.; Boyle, C.; DiMeglio, L. A.; Mastrandrea, L. D.; Abel, K.-L.; Cengiz, E.; Cemeroglu, P. A.; Aleppo, G.; Largay, J. F.; Foster, N. C.; et al. 2017. Evaluation of pump discontinuation and associated factors in the T1D exchange clinic registry. *Journal of diabetes science and technology*, 11(2): 224–232.

Xiu, L.; Chen, X.; Mao, L.; Zhang, E.; and Yu, G. 2024. Unveiling the influence of persuasion strategies on cognitive engagement: an ERPs study on attentional search. *Frontiers in Behavioral Neuroscience*, 18: 1302770.

Yao, Z.; and Yu, H. 2025. A survey on llm-based multi-agent ai hospital.

Yu, H.; Zhou, J.; Li, L.; Chen, S.; Gallifant, J.; Shi, A.; Li, X.; Hua, W.; Jin, M.; Chen, G.; et al. 2024. Aipatient: Simulating patients with ehrs and llm powered agentic workflow. *arXiv preprint arXiv:2409.18924*.

Zeng, Y.; Lin, H.; Zhang, J.; Yang, D.; Jia, R.; and Shi, W. 2024. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14322–14350.

Zhang, Z.; Yao, Z.; Zhou, H.; Yu, H.; et al. 2023. Ehrtutor: Enhancing patient understanding of discharge instructions. *arXiv preprint arXiv:2310.19212*.