ChatWise: A Strategy-Guided Chatbot for Enhancing Cognitive Support in Older Adults

Anonymous ACL submission

Abstract

Cognitive health in older adults presents a growing challenge. Although conversational interventions show feasibility in improving cognitive wellness, human caregiver resources remain overloaded. AI-based chatbots have shown promise, yet existing work is often limited to implicit strategies or heavily depends on training and label resources. In response, we propose a strategy-guided AI chatbot named ChatWise that follows a dual-level conversation reasoning framework. It integrates macro-level strategy planning and micro-level utterance generation to enable engaging, multi-turn dialogue tailored to older adults. Empirical results show that ChatWise closely aligns with professional human caregiver behaviors in offline evaluation using real clinic data, and achieves positive user cognitive and emotional responses in interactive simulations with digital twins, which significantly outperforms AI baselines that follow implicit conversation generation.

1 Introduction

The cognitive well-being of the elderly population is a pressing social concern, as evidenced by the prevalence of cognitive disorders within this population, often exacerbated by loneliness and isolation (Nicholson, 2012; Teo et al., 2023). According to WHO, approximately 14% of adults aged 60 and over experience mental health disorders, projected to affect 2.1 billion individuals by 2050 (Organization, 2024). Compared with other age groups, older adults are more vulnerable due to age-related changes in cognitive reserves (Salthouse, 2009) and reduced social connections (Nicholson, 2012; Teo et al., 2023). Such impact extends beyond older individuals to families and society, resulting in a reduced life quality and increased medical burden. In the meantime, modest delays in cognitive decline can significantly reduce dementia prevalence, and addressing social isolation could prevent 4% of dementia cases (Livingston et al., 2020). Interventions through guided conversations have shown efficacy in reducing loneliness and mitigating cognitive decline (Yu et al., 2021, 2022, 2023; Fiori and Jager, 2012). However, effective intervention requires sustained interaction and monitoring, which is limited by the availability of human companions, leading to inconsistent access or effectiveness.

Advances in artificial intelligence (AI) particularly Large Language Models (LLMs) have shown promise in augmenting human expertise with conversational support. Existing efforts range from sentiment-based chatbots (Liu et al., 2021; sen, 2024; Ryu et al., 2020; Tsai et al., 2021; You et al., 2023; Meyer et al., 2020) to audio assistants (Yang et al., 2024; Hong et al., 2024), ascribed to mature text-audio transformation techniques (Van Den Oord et al., 2016; Ren et al., 2019). However, these systems primarily default to interactions with implicit goals, which may fail to drive engaging conversations that necessitate strategic multi-turn interactions tailored to older adults. Some work has explored fine-tuning LLMs on domain-specific datasets to adapt to elderly care (Sun et al., 2025), which requires extensive labeled data and resources, while they may struggle to generalize across conversational contexts. In contrast, we take an orthogonal approach by focusing on the inherent reasoning capabilities of LLMs through in-context learning, which offers more resource efficiency and adaptability.

Our goal is to provide AI-powered, ***engaging*** *conversational support* for older adults that serves as an accessible complement to human companions, with the aspiration to improve their cognitive function and reduce social loneliness. Previous clinical studies aimed at socially isolated older adults showed the existence of causal relationships between interviewer strategy and interviewee response (Cao et al., 2021), which revealed that conversation behavior can have a *measurable* influence

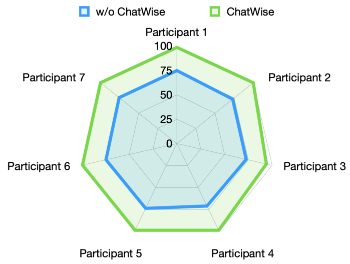


Figure 1: Alignment between conversational strategies of AI chatbots and human professionals across diverse participant data. Our design shows closer alignment with caregiver behavior in real conversation contexts compared to baseline chatbots. See Sec 4 for full results.

on interlocutors, and thus inspired us to develop principled yet efficient methods that transform clinical insights into dialogue design by leveraging the advanced reasoning ability of LLMs.

In response, we propose an LLM-driven chatbot named ChatWise. It employs *dual-layer* conversation generation that first derives categorized macro-level information to suggest meta-conversational strategies, which then guides the micro-level utterance generation to improve both user engagement and cognitive outcomes over multi-turn dialogue interactions. ChatWise is evaluated using both real, de-identified real-world dialogues between older adults and professional human caregivers of a clinical trial (I-CONECT, 2024), and synthetic interactive conversations using simulated users (*i.e.* digital twins) (Hong et al., 2024) modeled on such data.

Our work provides multifold contributions: (i) We introduce a clinically grounded chatbot that highly aligns with the behavior of professional human caregivers to older adults (Sec 4.1). Figure 1 overviews its alignment performance. It also empirically enhances simulated users' engagement and cognitive status, which significantly outperforms baseline chatbots (Sec 4.2). (ii) ChatWise follows a tuning-free, in-context learning framework for daily conversational support. Comparative studies demonstrated that providing macro-level strategies to guide conversation generation is the key contributor to enhancing user engagement. (iii) Its dual-level policy design can be readily applied to various LLMs, with consistent user cognitive gains across different backbone LLMs.

2 Related Work

AI-powered Chatbots for Older Adults have shown feasibility in preventing or detecting the cognitive issues to assist human professionals. Recent efforts span commercial products (sen, 2024; ell, 2024) and research prototypes focused on emotional support and audio assistance. (Sun et al., 2025) performed supervised fine-tuning on LLMs to enhance their performance on specialized nursing and elderly care tasks. Yang et al. (2024) introduced an LLM-based voice assistant designed to bridge communication between older adults and their healthcare providers. Liu et al. (2021) built an annotated dataset to tune LLMs for emotional support tasks. Ryu et al. (2020) developed a chatbot for the mental health of the elderly. Hong et al. (2024) constructed digital twins of the elderly. Unlike prior work, our approach prioritizes user conversational *engagement* through in-context reasoning, which has been empirically shown to enhance user cognitive status.

Dialogue Systems for Mental Health or Cognitive Stimulation: Recent studies have leveraged LLMs for the augmentation of emotional support conversations through *generated* dialogue data (Zheng et al., 2023; Jiang et al., 2023). Xygkou et al. (2024) assessed the acceptability of GPT-4-based conversational agents among people with dementia (PwD), and highlighted the importance of design considerations for this sensitive population. Similarly, Favela et al. (2023) implemented a conversational robot embedded with ChatGPTs for reminiscence therapy, with authors as simulated users, which showcased the potential of generative AI to support cognitive stimulation through interactive storytelling. Distinct from these works, our research introduces a dual-level dialogue approach combined with comprehensive evaluation methodologies to provide a finer-grained understanding of conversational engagement.

Conversational Strategies: Liu et al. (2021) curated a dataset with annotated strategies, demonstrating the effectiveness of *Helping Skills Theory* (Hill, 2020) in providing emotional support. Yuan et al. (2023) examined the causal relationships between dialogue acts (DAs) and participants' emotional states in a clinical trial (I-CONECT, 2024), emphasizing the impact of strategic interventions in tele-mediated dialogues. Seo et al. (2021) identified key strategies for improving child patient-provider communication through semi-structured interviews. However, few works have systematically integrated these strategies into the reasoning flow of AI chatbots. Our approach fills this gap by contextualizing structured conversational strategies for automatically enhancing user engagement.

Multi-Turn Chatbot Exploration: Recent advances in AI dialogue systems predominantly focused on short-turn interactions (Owan et al., 2023; Dam et al., 2024; Gao et al., 2025), with relatively limited attention to the challenges of multi-turn exchanges. While some pioneering research has explored multi-turn optimization through Reinforcement Learning (RL) approaches for LLMs (Verma et al., 2022; Zhou et al., 2024; Abdulhai et al., 2023; Gao et al., 2025), these methods were not specifically designed for supporting senior dialogue engagement. Our inference-based method introduces a strategy-compatible framework that aligns with the principles of RL-driven optimization, which can enable future extension to further enhance multi-turn conversation engagement.

3 ChatWise Design

ChatWise encompasses a hierarchical framework that integrates the insights of traditional clinical studies to drive LLM reasoning. It features a dual-level design with a *strategy provider* π_s that generates macro-level conversation strategies, which guides a *utterance generator* π_u to produce micro-level responses. In this paper, we interchangeably refer to ChatWise as the *interviewer* or the *moderator*, a role commonly defined in communication studies that guides and facilitates the conversation (Taboada, 2006),

For clarity, we denote a multi-turn conversation as $\tau = \{\mu_0, u_0, x_1, u_1, x_2, u_2, \cdots\}$ which comprises interviewee (user) utterances x_t and interviewer (ChatWise) utterances u_t at time step t, with a conversation starting point μ_0. $s_t = \{\mu_0, u_0, \cdots, x_t\}$ denotes the historical conversation up to timestep t. We then formulate ChatWise as a dual policy $\pi = \{\pi_s, \pi_u\}$, which maps the conversational *state* s_t to the next utterance u_t provided to the user, conditioned on meta information z given to the strategy provider π_s:

$$u_t \sim \pi(\cdot|s_t, z) \equiv \pi_u(\cdot|\pi_s(s_t|z)).$$

Below, we illustrate the process of this hierarchical conversation generation.

3.1 Strategy Pool

To first develop clinically grounded strategies, we establish a strategy pool containing dialogue acts (DA) curated from clinical intervention studies involving older adults. These DAs, such as *asking an open question* (*e.g.* "What is your favorite movie") or *showing empathy* (*e.g.*"That sounds really meaningful"), serve as macro-level actions and atomic communicative units that convey distinct conversational intentions (Searle, 1976), which have also shown causal relationships with participant emotions in conversations (Cao et al., 2021). Specifically, we extract DAs that exist in conversations from a real tele-health clinical trial (Yuan et al., 2023), which represents a subset of the complete DAMSL taxonomy (Allen and Core, 1997; Searle, 1976). We further augment this DA with strategies from prior emotional support dialogue dataset (Liu et al., 2021) to form a comprehensive strategy pool $\mathcal{A}$. We then treat each strategy serves as a macro-level action candidate $a \in \mathcal{A}$ and refine it with a *definition* and an *example* to serve as in-context information z for enhancing the reasoning of strategy π_s (See Appendix A.5 for a complete DA set).

3.2 Emotion Annotation

To enhance context-awareness during conversations and support post-conversation analysis, the strategy provider π_s is additionally prompted to infer the user's current emotion based on historical conversations s_t. At timestep t of the conversation, π_s will categorize user emotion into one of the five classes: *joy, neutral, sadness, anger, and surprise*, given a calibrated system prompt. Our comparative study shows that ChatWise can lead to detectable positive emotion changes over multi-turn conversations, which aligns with conversations provided by human professionals (Sec 4.5).

3.3 From Strategy to Utterance Generation

The strategy provider π_s processes the dialogue history s_t, contextualized by the strategy pool information z to decide the most appropriate strategies for continuing a conversation. We limit each interviewer's utterance to contain one or two strategies, *i.e.* $\mathbf{a}_t = \{a_t^i\}_{i \le 2} \sim \pi_s(\cdot|s_t, z)$. Our rationale follows counseling studies (Zhang and Danescu-Niculescu-Mizil, 2020) that dialogue intentions can be *forward*, *e.g.* initiating new topics via an open question, or *backward*, *e.g.* responding with acknowledgment, or both to transit between intentions or topics. Thus, π_s is prompted to comply with one of the three conditions: (i) a forward strategy, (ii) a backward strategy, or (iii) a backward strategy followed by a forward strategy.

The selected strategies $\mathbf{a}_t$, and *optionally*, the user's current emotion label e_t, and the conversation history s_t are used as inputs to the utterance

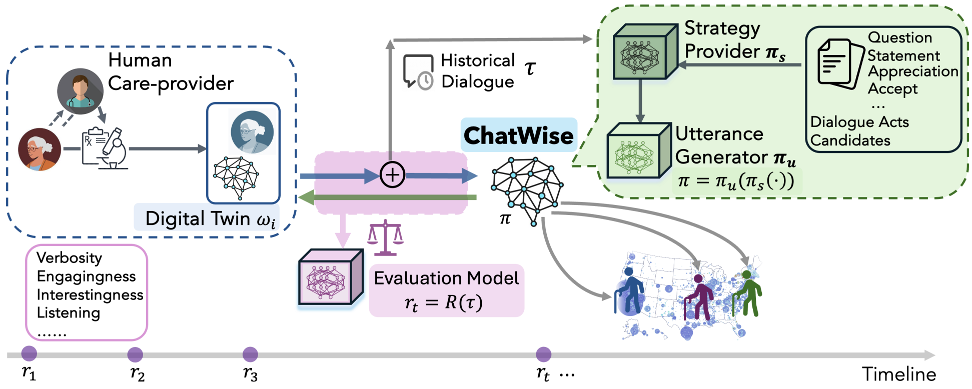


Figure 2: ChatWise employs a dual-level policy design that contextualizes a clinically-derived *strategy pool* to generate *macro-level* actions, which then guide *micro-level* utterance generation. It is iteratively developed using digital twins as simulated users. Prior to a future real user study, real and simulated data analysis show that ChatWise closely aligns with professional human caregivers' behavior, leading to cognitive gains of simulated users accumulated over multi-turn interactions.

generator π_u. Our ablation study showed that providing meta-level strategies has a dominant effect over providing emotion information alone for conversational engagement (Sec 4.3). To ensure natural flow, π_u will first improvise a few rounds as warm-ups before adopting the suggested strategies. We also draw on clinical guidance to let π_u encourage users to choose topics rather than imposing them. Figure 2 overviews this design.

4 Experiments

We conducted various experiments focusing on answering the following questions:

Q1. Does ChatWise's behavior align with human professionals when responding to real conversation context sampled from clinic trials?
Q2. How does ChatWise compare to baseline AI alternatives in terms of multifaceted conversational engagement metrics?
Q3. How do factors such as the conversation turns and different user characteristics influence ChatWise's performance?

To address these questions, we evaluated ChatWise on two complementary scenarios: we first conducted ***offline*** evaluations using data from a real clinical trial (Sec 4.1), then we performed ***interactive*** experiments, in which conversations are stochastically generated between ChatWise and digital twins as simulated users (Sec 4.2). Below, we summarize the configurations, metrics, and main results of each experimental setting and defer more details to the Appendix.

4.1 Offline Evaluation on Real Clinical Data

4.1.1 Data Preprocessing

We extracted dialogues from the I-CONECT clinic trials (I-CONECT, 2024), with each in text format converted from video conversations. Short dialogues below 40 *turns* were excluded, where a *turn* represents an uninterrupted, continuous utterance by a single speaker. Two dialogue sets were subsampled and de-identified: (1) 150 randomly sampled dialogues to assess the overall strategy alignment, and (2) one dialogue per week from 7 randomly selected participants to assess ChatWise's alignment robustness to different participants and time periods over the clinical trial.

4.1.2 Offline Evaluation Metric

We define Strategy Match Percentage (SMP) to quantify the alignment between the conversational strategies generated by a chatbot and those generated by a human professional caregiver from the I-CONECT clinic data. Let $\mathbf{a}_t$ denote the set of caregiver-provided strategies annotated from a real conversation τ at turn (timestep) t, and $\mathbf{a}'_t$ the strategies from a simulated moderator (ChatWise or utterance generator), conditioned on the same real context s_t. Let $\mathbb{I}(a'_t \in \mathbf{a}_t)$ be an indicator function that returns 1 if a chatbot-generated strategy a'_t matches one of the human strategies $\mathbf{a}_t$, and 0 otherwise. The SMP is computed as following:

$$\mathrm{SMP}_t(\mathbf{a}_t, \mathbf{a}'_t) = \frac{1}{|\mathbf{a}_t|} \sum_{a'_t \in \mathbf{a}'_t} \mathbb{I}(a'_t \in \mathbf{a}_t).$$

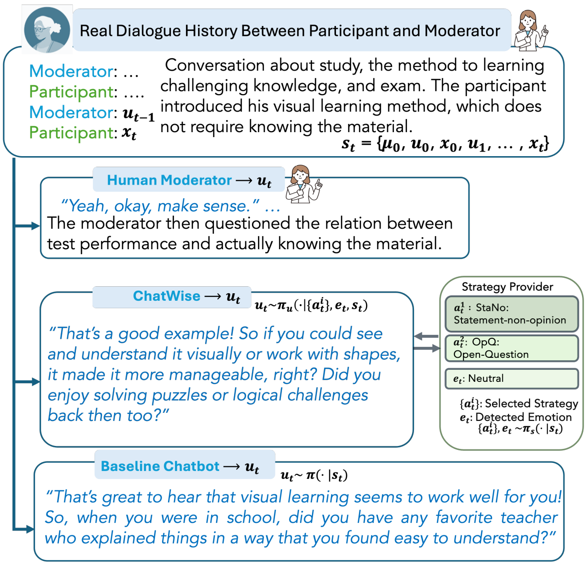


Figure 3: Comparison of moderator strategies: Given real dialogue contexts, responses from a baseline (generated), ChatWise (generated), and the human moderator (real) are compared *w.r.t.* underlying strategies. A real sample is shown here while omitting participant data.

4.1.3 Strategy Label Annotation

Given a real dialogue denoting as $\tau = \{\mu_0, u_0, x_1, u_1, x_2, u_2, \cdots\}$ with caregiver (moderator) utterance u_t and participant utterance x_t at turn t, we leveraged the macro-level strategy definition in Sec 3.1 to prompt an LLM as annotator (GPT-4o) and detect strategies $\mathbf{a}_t$ behind u_t as the golden label. Denoting $s_t = \{\mu_0, u_0, x_1, \cdots, x_t\}$ as the real conversation context, we feed s_t to a ***baseline*** chatbot to get utterance action $\tilde{u}_t$. The baseline is identical to the utterance generator π_u of ChatWise yet does not receive strategy guidance from π_s. Similarly, we get strategy labels $\tilde{\mathbf{a}}_t$ behind $\tilde{u}_t$ using the above annotator. We then input s_t to ChatWise's strategy provider π_s to get proposed strategy $\mathbf{a}'_t \sim \pi_s(s_t)$. We report the strategy alignment of $\text{SMP}_t(\mathbf{a}_t, \mathbf{a}'_t)$ (w/ ChatWise) and $\text{SMP}_t(\mathbf{a}_t, \tilde{\mathbf{a}}_t)$ (w/o ChatWise) over different turns t and average the result across dialogues. Figure 3 illustrates the process of collecting each moderator's response based on real conversation history.

4.1.4 Overall Strategy Alignment

As shown in Figure 5, which averages results over 150 dialogues, ChatWise consistently achieves an SMP close to 1.0, indicating a strong alignment with the strategies employed by real caregivers throughout the conversation. In contrast, the baseline chatbot deviates from human behavior when lacking action guidance. Both chatbots used GPT-4o as the utterance generator π_u, and o3-mini was used for π_s in ChatWise. We do not employ micro-level utterance similarity, as the I-CONECT interventions involve spontaneous daily conversations, rather than fixed question answering, unlike domains in programming or math. Whereas, on a macro level, ChatWise mostly mirrors the strategic choices of human professionals who are well trained to follow clinical protocols.

To illustrate how our method behaves differently from the baseline, we collect the real dialogue context and different moderator responses where $\mathbf{a_t} = \mathbf{a}'_t \neq \tilde{\mathbf{a}}_t$. and show the discrepancy heatmap. Raw dialogue content is omitted to comply with the data usage agreement. Results in Figure 4 demonstrate that while "*open question*" is a commonly dominant strategy (forward DA) for both human and chatbot moderators, the primary strategy difference lies in how they lead into the question (*e.g.* "*acknowledge*", "*restatement*"), which may in turn influence the specific question ultimately asked.

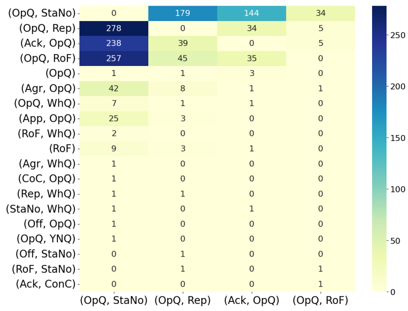


Figure 4: Strategy distribution for dialogue samples where $\mathbf{a_t} = \mathbf{a}'_t \neq \tilde{\mathbf{a}}_t$. The vertical axis indicates ground-truth strategies $\mathbf{a_t}$ extracted from human professional utterances, and the horizontal axis denotes strategies $\tilde{\mathbf{a}}_t$ detected from the baseline chatbot.

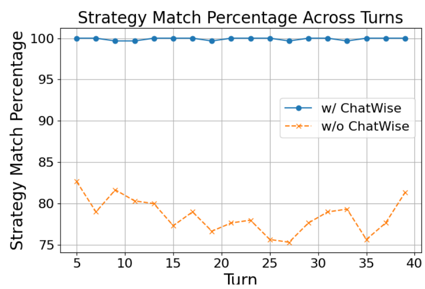


Figure 5: SMP between human caregiver and chatbots on real conversation data, where w/o ChatWise denotes the baseline without strategy guidance. Our method strongly aligns with human caregiver behavior.

Robustness to Participant Heterogeneity: Figure 6 presents the SMP of ChatWise across dialogue turns among participants. ChatWise consistently outperforms the baseline and maintains notably higher SMP. These results demonstrate the robust alignment of ChatWise with human caregiver strategies across diverse participants and their interaction contexts, which implies its potential for personalized dialogue systems.

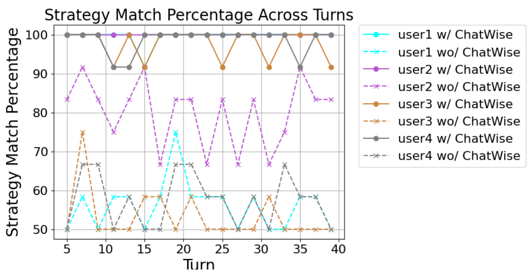


Figure 6: SMP broken down by participants over 40 conversation turns, where each color indicates a specific person. Solid lines represent ChatWise, and dashed lines denote the baseline. ChatWise consistently aligns to human caregivers given different participant contexts.

Alignment Consistency Over Timeline: Figure 7 presents the SMP over time, where dialogues were sampled from different weeks over the clinical intervention. ChatWise maintains consistently high SMP (near 1.0) across all timesteps, where the baseline shows lower and unstable performance. This indicates ChatWise's temporal robustness and its potential to facilitate the development of long-term cognitively supportive dialogues.

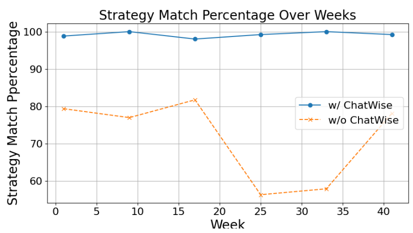


Figure 7: Strategy alignment of ChatWise is consistent as real conversations progress over weeks.

4.2 Evaluation with Digital Twins

4.2.1 Conversation Generation

While evaluation on real clinic data shows strong alignment between ChatWise and professional caregivers, the conversations are fixed, thus lacking indications whether and how ChatWise may influence user cognitive or emotional status during conversations. To enable controlled and responsive development of ChatWise prior to future human user studies, we used 9 digital twins provided by Hong et al. (2024) as simulated users with different personas, which are fine-tuned LLMs calibrated to mimic the linguistic behaviors of older adults, some of whom show MCI symptoms. All digital twins are trained using *real, de-identified* dialogue data from the I-CONECT clinical trial. We collected 20 trajectories of conversations between ChatWise (moderator) and each digital twin as a user, with each conversation containing 20 turns. The same settings were applied to the baseline chatbot for fair comparisons.

4.2.2 Metrics for Interactive Engagement

User Verbosity: Given a conversation sequence τ that consists of user utterance x and moderator utterance u, we primarily measure user engagement through the user's talkativeness compared to the moderator, *i.e.* the user's verbosity (v), defined as $v = \frac{\sum_{x\in\tau}|x|}{\sum_{u\in\tau}|u|}$, where $|x|$ ($|u|$) denotes the number of tokens in an utterance x (u). This metric follows a clinical study (Yu et al., 2021) that suggests reducing the moderator talkativeness while encouraging participant (user) expression.

Cognitive Win Rate: See et al. (2019) defined different aspects for evaluating conversation quality, from which we select 3 that are focused on assessing user cognitive status: *Listening*, *Fluency*, and *Making Sense*. All three metrics evaluate the ***user*** behavior rather than the overall dialogue quality considering both user and moderator behaviors.

Instead of providing a numerical score for each cognitive metric that might be unstable, we adopt the ***Win Rate (WR)*** definition from prior work (Rafailov et al., 2023) to compare pairs of dialogues generated with and without ChatWise. Dialogue pairs are randomly matched so long as they are collected from the same digital twin (simulated user) for evaluation. WR is defined as the proportion of pairs in which the dialogue generated with ChatWise is preferred over the baseline using an LLM-as-judge, based on a given cognitive metric.

To mitigate the LLM's *position bias* issues observed in prior work (Shi et al., 2024; Yu et al., 2024), we excluded sample pairs with inconsistent preference labels when the order of the two samples was reversed and calculated WRs using only the remaining consistent pairs.

4.2.3 Models and Baseline

We evaluated ChatWise with 3 different LLM backbones as strategy providers: GPT-4o, o3-mini, and Llama3.1-405B. For each setting, a baseline adopts the same utterance generator as ChatWise while without receiving strategy guidance. The WR of *baseline* is re-

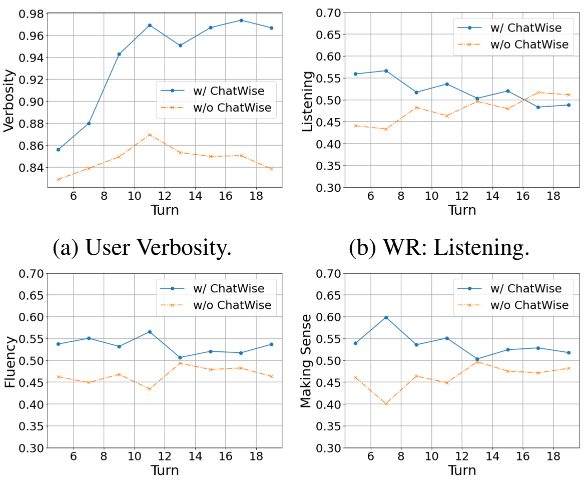
Strategy Provider	Verbosity ↑	Listening ↑	Fluency ↑	Making Sense ↑
baseline	0.7398	**0.5249**	0.4986	0.5024
GPT-4o	0.8635	0.4962	0.4748	0.4786
o3-mini	**0.8643**	0.4884	**0.5368**	**0.5180**
Llama 3.1-405B	0.8083	0.4407	0.4926	0.4963

Table 1: Multifaceted evaluations on cognitive engagement with data sampled by interacting with digital twins over 20 turns, using different LLMs serving as strategy providers, where *Listening*, *Fluency*, and *Making Sense* metrics report win rates, and baseline denotes a conversation generator without receiving strategy guidance. All evaluations focus on *user* reactions.

ported as the complement of ChatWise's average WR: $\text{WR}_{\text{baseline}} = 1 - \text{average}(\text{WR}_{GPT-4o} + \text{WR}_{o3-mini} + \text{WR}_{Llama3.1-405B})$, where WR_X indicates the WR of ChatWise, given model X as the strategy provider.

4.2.4 Performance Overview

Effectiveness of Strategy-Guided Generation: As shown in Table 1, our design mostly demonstrated better user engagement compared to baselines across tested LLMs as the strategy provider. Particularly, the best-performing model in our setting, o3-mini, was optimized for STEM reasoning tasks, which indicates that strong reasoning ability can also benefit inferring dialogue strategies.



(a) User Verbosity. (b) WR: Listening.

(c) WR: User Fluency. (d) WR: Making sense.

Figure 8: Multifaceted Engagement Evaluation across conversation turns, excluding the first two warm-up turns. ChatWise's performance gain over the baseline exists over multi-turn dialogues.

Performance with Dialogue Progression: We analyzed ChatWise's performance over increasing dialogue length by truncating conversations at turn t and computing metrics accordingly, as shown in Figure 8. Results revealed a clear upward trend in *verbosity*, which indicates increased talkativeness of the user rather than the ChatWise, as opposed to the baseline. Meanwhile, the WR gains of ChatWise either remain stable or show a slight decline as conversations progress, which we infer are subject to the length bias of LLMs (Dubois et al., 2024) as the overall number of tokens of ChatWise-generated dialogues tends to decrease in later turns.

4.3 Ablation Study

We evaluated ChatWise with and without user *emotion* in the strategy provider's output. As shown in Table 2, where both strategy providers used GPT-4o as the backbone, removing emotion information led to only a slight drop of performance, while removing ChatWise entirely resulted in a much lower verbosity score, which highlights that strategy guidance itself is the key driver of user engagement.

Method	ChatWise	ChatWise w/o emotion	w/o ChatWise
Verbosity	0.8635	0.8463	0.7398

Table 2: Ablative study on the performance of user verbosity when removing different contextualized information. w/o emotion denotes the strategy provider without user emotion information as input.

4.4 ChatWise's Robustness to User Persona

As shown in Figure 9, the gain of ChatWise interacting with 9 simulated users is consistently significant, in which the metrics were log-normalized. This indicates the robustness of ChatWise given varying senior characteristics.

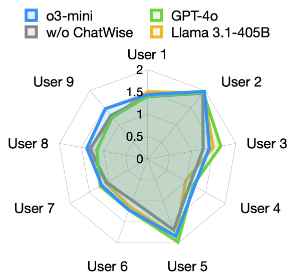


Figure 9: User *verbosity* log-normalized across 9 digital twins. Given varying strategy provider backbones, our design consistently outperforms the baseline (w/o ChatWise) with non-strategic conversations.

4.5 Analysis of User Emotion Transitions

To analyze if ChatWise can support positive emotional shifts during conversations, we selected four digital twins and collected 40 additional dialogues from each for deeper analysis, with the main findings summarized below.

Transient user emotions: We define an emotion transition triplet as $(\mathrm{e}_t, u_t, \mathrm{e}_{t+1})$, where e_t (e_{t+1}) represents the user's emotion at turn t $(t+1)$, and u_t is the moderator utterance in between. We computed the average occurrence of emotion triplets and showed the top 15 most frequent in Figure 10. Most triplets reflect unchanged user emotions, which indicates difficulty in either influencing or detecting emotional shifts within a short turn.

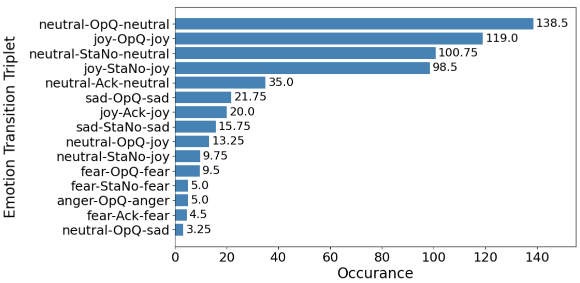
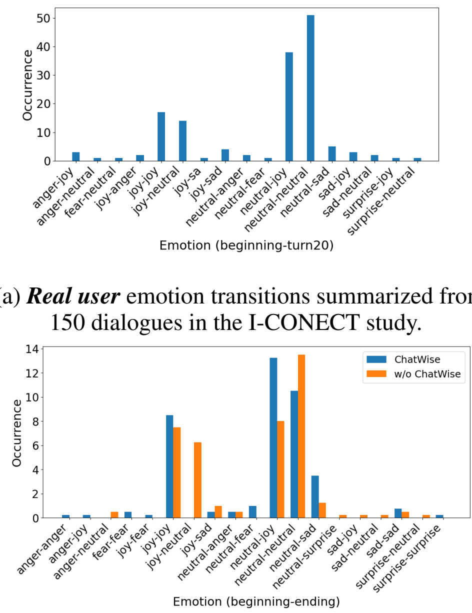


Figure 10: Average occurrence of each emotion transition triplet across the samples of each digital twin.



(a) ***Real user*** emotion transitions summarized from 150 dialogues in the I-CONECT study.

(b) ***Digital twin*** emotion transitions summarized over 80 dialgoues.

Figure 11: User emotion shifts in conversations between digital-twin and ChatWise (Figure 11b) align well with real-user emotion shifts in the clinical study (Figure 11a). More *neutural-to-joy* transitions were observed given ChatWise compared to the baseline.

User Emotions over Multi-Turn Conversations: We then analyzed user emotion changes from the beginning to the end of dialogues and calculated averaged occurrence across digital twins (Figure 11b). Over 48% users experienced emotional shifts post-dialogue, with significantly more positive changes after engaging with ChatWise. This implies that while transient emotions are difficult to track, strategic conversational support may improve user emotions over multi-turn conversations.

Predominant Strategies: We identified the top 10 most frequent strategies across digital twins (Figure 13) and found that predominant interaction strategies remained consistent across user types (See Appendix A.8). Particularly, *Open Question*, *Statement-non-opinion*, and *Acknowledgment* strategies dominate ChatWise driven conversations.

5 Conclusion and Future Work

We present a dual-level framework for AI chatbots that supports multi-turn conversations for older adults by integrating clinical insights into LLM reasoning. Extensive evaluations based on both real clinical data and generated conversations showed that our method aligns well with the behavior of professional human caregivers while robustly enhancing the cognitive status of older adults through simulation studies. Our future work includes training-based methods for further optimizing multi-turn interaction experience and investigating the acceptability and feasibility of our design through real user studies.

6 Ethical Considerations

This study involved simulated dialogues using digital twins trained on *de-identified* conversation data from a clinical trial involving older adults. Offline analysis of this dataset was conducted to evaluate the similarity between chatbot and caregiver strategies. This project did not use any personally identifiable information, and all data processing followed institutional privacy and research ethics guidelines. Potential risks are controllable, which include privacy concerns of emotion detection and misinterpretation of conversation due to AI-based evaluation. To mitigate this, the emotional detection module in the system is made optional and can be omitted as configured. Win rate calculations are based on a sanity check where we omitted pair samples that receive inconsistent preference labels by LLM-as-the-judge after swapping their positions. We have manually cross-validated a sufficient subset of AI-generated analyses to reduce bias. A real-user study is planned for future validation and will undergo full Institutional Review Board (IRB) approval.

Limitations

All digital twins were provided as fine-tuned GPT-3.5 APIs, and their high economic cost, at the time when this project was going on, constrained the volume of dialogues we generated. To mitigate this, future work will explore accessing digital twins using LLaMA 3.1 to reduce sampling costs while maintaining conversational quality. LLM evaluations have been discussed to show position bias. Regarding win rate calculation, after we took mitigation methods and filtered out pairs with inconsistent preference labels, the remaining samples contributed to 81% of the data before filtering. Future mitigation methods will include learning a reward function for more fine-grained and robust evaluation. Furthermore, our study relies either on offline static data or simulated dialogues rather than dynamic user interactions. A real user study is needed to validate the system's acceptability and feasibility in real-world settings, which we plan to incorporate into our future research.

References

2024. Senior talk: Ai chatbot for mental health support for seniors.

2024. Your ai sidekick for happier, healthier aging.

Marwa Abdulhai, Isadora White, Charlie Snell, Charles Sun, Joey Hong, Yuexiang Zhai, Kelvin Xu, and Sergey Levine. 2023. Lmrl gym: Benchmarks for multi-turn reinforcement learning with language models. *arXiv preprint arXiv:2311.18232*.

James Allen and Mark Core. 1997. Draft of damsl: Dialog act markup in several layers.

Shuyi Cao, Lizhen Qu, and Leimin Tian. 2021. Causal relationships between emotions and dialog acts. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

Sumit Kumar Dam, Choong Seon Hong, Yu Qiao, and Chaoning Zhang. 2024. A complete survey on llm-based ai chatbots. *arXiv preprint arXiv:2406.16937*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Jesus Favela, Dagoberto Cruz-Sandoval, and Mario O. Parra. 2023. Conversational agents for dementia using large language models. In *2023 Mexican International Conference on Computer Science (ENC)*, pages 1–7.

Katherine L Fiori and Justin Jager. 2012. The impact of social support networks on mental and physical health in the transition to older adulthood: A longitudinal, pattern-centered approach. *International Journal of Behavioral Development*, 36(2):117–129.

Zhaolin Gao, Wenhao Zhan, Jonathan Daniel Chang, Gokul Swamy, Kianté Brantley, Jason D. Lee, and Wen Sun. 2025. Regressing the relative future: Efficient policy optimization for multi-turn RLHF. In *The Thirteenth International Conference on Learning Representations*.

Clara E Hill. 2020. *Helping skills: Facilitating exploration, insight, and action*. American Psychological Association.

Junyuan Hong, Wenqing Zheng, Han Meng, Siqi Liang, Anqing Chen, Hiroko H Dodge, Jiayu Zhou, and Zhangyang Wang. 2024. A-conect: Designing ai-based conversational chatbot for early dementia intervention. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

I-CONECT. 2024. Study details.

Jiyue Jiang, Sheng Wang, Qintong Li, Lingpeng Kong, and Chuan Wu. 2023. A cognitive stimulation dialogue system with multi-source knowledge fusion for elders with cognitive impairment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10628–10640, Toronto, Canada. Association for Computational Linguistics.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

Gill Livingston, Jonathan Huntley, Andrew Sommerlad, David Ames, Clive Ballard, Sube Banerjee, Carol Brayne, Alistair Burns, Jiska Cohen-Mansfield, Claudia Cooper, et al. 2020. Dementia prevention, intervention, and care: 2020 report of the lancet commission. *The lancet*, 396(10248):413–446.

Ashley ND Meyer, Traber D Giardina, Christiane Spitzmueller, Umber Shahid, Taylor MT Scott, and Hardeep Singh. 2020. Patient perspectives on the usefulness of an artificial intelligence–assisted symptom checker: cross-sectional survey study. *Journal of medical Internet research*, 22(1):e14679.

Nicholas R Nicholson. 2012. A review of social isolation: an important but underassessed condition in older adults. *The journal of primary prevention*, 33:137–152.

World Health Organization. 2024. Ageing and health.

Risako Owan, Libby Ferland, Naome Etori, Wilma Koutstaal, and Maria Gini. 2023. Conversational agents for elderly users. In *Workshop Assistive Robotics for Citizens at IROS*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems*, 32.

Hyeyoung Ryu, Soyeon Kim, Dain Kim, Sooan Han, Keeheon Lee, and Younah Kang. 2020. Simple and steady interactions win the healthy mentality: designing a chatbot service for the elderly. *Proceedings of the ACM on human-computer interaction*, 4(CSCW2):1–25.

Timothy A Salthouse. 2009. When does age-related cognitive decline begin? *Neurobiology of aging*, 30(4):507–514.

John R Searle. 1976. A classification of illocutionary acts1. *Language in society*, 5(1):1–23.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1702–1723, Minneapolis, Minnesota. Association for Computational Linguistics.

Woosuk Seo, Ayse G Buyuktur, Sanya Verma, Hyeryoung Kim, Sung Won Choi, Laura Sedig, and Sun Young Park. 2021. Learning from healthcare providers' strategies: Designing technology to support effective child patient-provider communication. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Lin Shi, Chiyu Ma, Wenhua Liang, Weicheng Ma, and Soroush Vosoughi. 2024. Judging the judges: A systematic investigation of position bias in pairwise comparative assessments by llms. *arXiv preprint arXiv:2406.07791*.

Qiao Sun, Jiexin Xie, Nanyang Ye, Qinying Gu, and Shijie Guo. 2025. Enhancing nursing and elderly care with large language models: An ai-driven framework. *Proceedings of the 31st International Conference on Computational Linguistics*.

Maite Taboada. 2006. Spontaneous and non-spontaneous turn-taking. *Pragmatics*, page 329–360.

Rui Hong Teo, Wen Hui Cheng, Ling Jie Cheng, Ying Lau, and Siew Tiang Lau. 2023. Global prevalence of social isolation among community-dwelling older adults: a systematic review and meta-analysis. *Archives of gerontology and geriatrics*, 107:104904.

Chun-Hua Tsai, Yue You, Xinning Gui, Yubo Kou, and John M Carroll. 2021. Exploring and promoting diagnostic transparency and explainability in online symptom checkers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, et al. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 12.

Siddharth Verma, Justin Fu, Mengjiao Yang, and Sergey Levine. 2022. Chai: A chatbot ai for task-oriented dialogue with offline reinforcement learning. *arXiv preprint arXiv:2204.08426*.

Anna Xygkou, Chee Siang Ang, Panote Siriaraya, Jonasz Piotr Kopecki, Alexandra Covaci, Eiman Kanjo, and Wan-Jou She. 2024. Mindtalker: Navigating the complexities of ai-enhanced social engagement for people with early-stage dementia. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

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. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(2):1–35.

Yue You, Chun-Hua Tsai, Yao Li, Fenglong Ma, Christopher Heron, and Xinning Gui. 2023. Beyond self-diagnosis: how a chatbot-based symptom checker should respond. *ACM Transactions on Computer-Human Interaction*, 30(4):1–44.

Kexin Yu, Benjamin M Hampstead, Katherine Wild, Lisa C Silbert, and Hiroko H Dodge. 2022. Examining i-conect intervention effect on psychosocial wellbeing. *Alzheimer's & Dementia*, 18:e059654.

Kexin Yu, Katherine Wild, Kathleen Potempa, Benjamin M Hampstead, Peter A Lichtenberg, Laura M Struble, Patrick Pruitt, Elena L Alfaro, Jacob Lindsley, Mattie MacDonald, et al. 2021. The internet-based conversational engagement clinical trial (i-conect) in socially isolated adults 75+ years old: randomized controlled trial protocol and covid-19 related study modifications. *Frontiers in digital health*, 3:714813.

Kexin Yu, Chao-Yi Wu, Lisa C Silbert, Jeffrey A Kaye, and Hiroko H Dodge. 2023. I-conect intervention effects on weekly time spent outside of home and social contacts among socially isolated older adults with normal cognition and mild cognitive impairment. *Alzheimer's & Dementia*, 19:e077984.

Yijiong Yu, Huiqiang Jiang, Xufang Luo, Qianhui Wu, Chin-Yew Lin, Dongsheng Li, Yuqing Yang, Yongfeng Huang, and Lili Qiu. 2024. Mitigate position bias in large language models via scaling a single dimension. *arXiv preprint arXiv:2406.02536*.

Fengpei Yuan, Wenjun Zhou, Hiroko H. Dodge, and Xiaopeng Zhao. 2023. Short: Causal structural learning of conversational engagement for socially isolated older adults. *Smart Health*, 28:100384.

Justine Zhang and Cristian Danescu-Niculescu-Mizil. 2020. Balancing objectives in counseling conversations: Advancing forwards or looking backwards. *arXiv preprint arXiv:2005.04245*.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023. AugESC: Dialogue augmentation with large language models for emotional support conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568, Toronto, Canada. Association for Computational Linguistics.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. 2024. Archer: Training language model agents via hierarchical multi-turn rl. *arXiv preprint arXiv:2402.19446*.

A Appendix

A.1 Prompt design

The prompt used in the experiments and structured output design are available at https://anonymous.4open.science/r/ChatWise-8F53, including:

- Strategy provider system prompt.
- Moderator initial system prompt.
- Moderator system prompt with strategies.
- Strategy provider system prompt for ablation study.
- Moderator system prompt with strategies for ablation study.
- Structured output class for OpenAI models as strategy provider.
- Structured output class for OpenAI models as strategy provider in ablation study.
- System prompt for GPT-4o to extract the strategy given by Llama3.1.

A.2 Offline Data Preparation

We extracted dialogue content from the I-CONECT video conversations and processed it into OpenAI-compatible dialogue history format. Dialogues with fewer than 40 turns were excluded. We then subsampled two subsets for evaluation purposes. To assess the overall strategy alignment of ChatWise, we randomly sampled 150 dialogues. To evaluate the robustness of ChatWise across different participants and time periods, we randomly selected 7 participants who completed the full study and sampled one dialogue from each of their sessions during weeks 1, 9, 17, 25, 33, and 41. All user identifiers and personally identifiable information have been anonymized.

The OpenAI-compatible dialogue history format example:[{"role": "system", "content": system_prompt}, {"role": "user", "content": user_content1}, {"role": "assistant", "content":assistant_content1}, {"role": "user", "content": user_content2}, {"role": "assistant", "content": assistant_content2}, ...].

A.3 Data Generation Configuration

By default, we used o3-mini as strategy provider. GPT-4o serves as an utterance generator in all settings. W/o ChatWise denotes the baseline, which is using GPT-4o as the utterance generator only, without a strategy provider. We tested different LLM backbones as strategy providers in ChatWise. Considering the Llama3.1-405B does not support structured output, we applied GPT-4o as the strategy extractor to structure the strategy output of Llama3.1-405B. We employed GPT-4o as the judge to select the preferred one when computing the Win Rate. The prompt for selecting the preferred response is listed in Figure 12. The following are the configurations of each model:

Utterance generator (GPT-4o):
n=1
max_tokens=1024
top_p=1
temperature=1

GPT-4o, o3-mini as strategy provider:
n=1
max_tokens=1024
top_p=1
temperature=1
response_format=Strategy

Llama3.1-405B as strategy provider:
top_p: 0.9
max_tokens: 1024
temperature: 0.6
presence_penalty: 0
frequency_penalty: 0

GPT-4o as strategy extractor:
n=1
max_tokens=1024
top_p=1
temperature=1

GPT-4o as judge:
n=1
max_tokens=1024
top_p=1
temperature=1

A.4 Win rate for w/o ChatWise

The Win rate for w/o ChatWise is defined as:

$$1 - \text{average}(\text{WR}_{GPT-4o} + \text{WR}_{o3-mini} + \text{WR}_{Llama3.1-405B})$$

,where WR_X is ChatWise's Win rate against the baseline with X as strategy provider.

A.5 Dialogue Acts

The map of strategy to its corresponding abbreviated tag is listed in Table3.

There are two kinds of strategies: backward-looking and forward-looking. Backward-looking

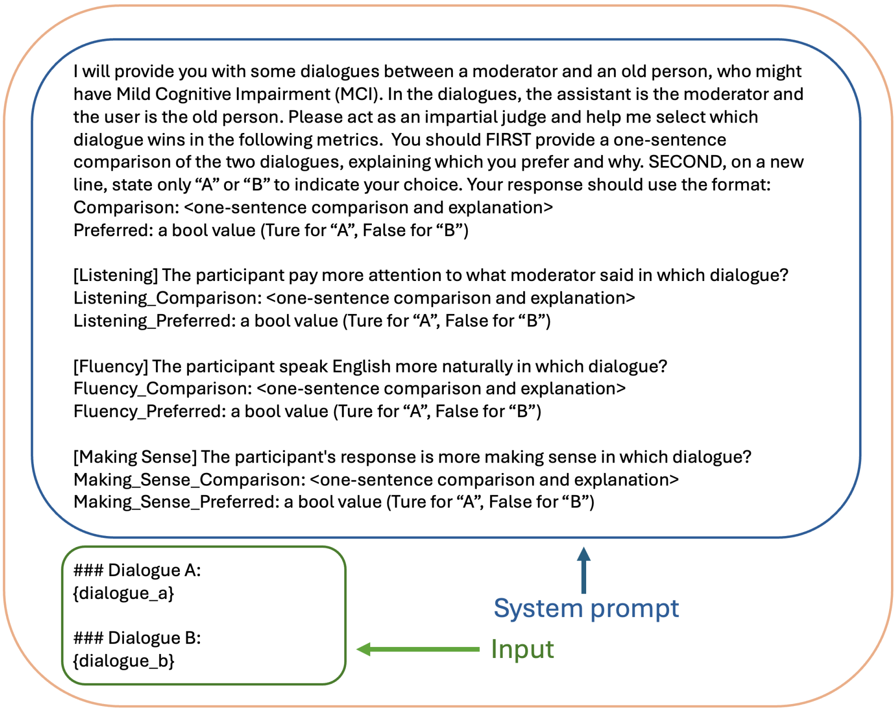


Figure 12: Prompt for GPT-4o as a Judge.

strategies reflect how the current utterance relates to the previous discourse. Forward-looking strategies reflect the current utterance constrains the future beliefs and actions of the participants and affects the discourse. Table 4 and Table 5 provide definitions and examples of each.

A.6 Log-normalization

The following is the log-normalization function, where y is the normalized result, x is the input variable.

$$y = \ln(4x + 1)$$

A.7 Primary Strategies

We calculated the average occurrence of each strategy across each digital twin and listed their top 10 most frequently occurring strategies, as shown in Figure 13.

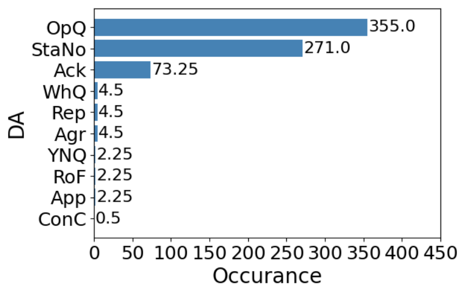


Figure 13: Strategy occurrence across digital twins.

Strategy	Tag
Acknowledge (Backchannel)	Ack
Statement-non-opinion	StaNo
Statement-opinion	Sta
Affirmation and Reassurance	Agr
Appreciation	App
Conventional-closing	ConC
Hedge	H
Other	Oth
Quotation	Quo
Action-directive	AcD
Collaborative Completion	CoC
Restatement or Paraphrasing	Rep
Offers Options Commits	Off
Self-talk	Sel
Apology	Apo
Reflection of Feelings	RoF
Yes-No-Question	YNQ
Wh-Question	WhQ
Declarative Yes-No-Question	DYNQ
Open-Question	OpQ
Or-Clause	OrC
Conventional-opening	CoO
Self-disclosure	Sd
Providing Suggestions	PS
Information	I

Table 3: Strategy to its corresponding tag. The strategies are drawn from the DAs in DAMSL (Allen and Core, 1997) that are used by telehealth clinical trials (Yuan et al., 2023), integrated strategies from prior emotional support dataset (Liu et al., 2021) .

Strategy	Definiton	Example
StaNo	A factual statement or descriptive utterance that does not include an opinion.	Me, I'm in the legal department.
Ack	A brief utterance that signals understanding, agreement, or active listening.	Uh-huh.
Sta	A statement that conveys a personal belief, judgment, or opinion.	I think it's great
Agr	Affirm the help seeker's strengths, motivation, and capabilities and provide reassurance and encouragement.	That's exactly it.
App	An expression of gratitude, admiration, or acknowledgment of another's effort or input.	I can imagine.
ConC	A formal or socially standard utterance signaling the end of a conversation.	Well, it's been nice talking to you.
H	An expression that introduces uncertainty or qualification to a statement, often to soften its impact.	I don't know if I'm making any sense or not.
Oth	Exchange pleasantries and use other support strategies that do not fall into the above categories.	Well give me a break, you know.
Quo	A direct or indirect repetition of someone else's words.	Albert Einstein once said, "Imagination is more important than knowledge."
AcD	A command, request, or suggestion directing someone to take action.	Why don't you go first
CoC	A continuation or completion of someone else's utterance in a collaborative manner.	If we want to make it to the top of the mountain before sunset, we should...
Rep	A simple, more concise rephrasing of the help-seeker's statements that could help them see their situation more clearly.	It sounds like you're saying that you're struggling to stay on top of your work, and it's leaving you feeling overwhelmed.
Off	A statement proposing choices, making a commitment, or offering to do something.	I'll have to check that out
Sel	An utterance directed at oneself, often reflecting internal thought processes or problem-solving.	What's the word I'm looking for
Apo	An expression of regret or asking for forgiveness.	I'm sorry.
RoF	Articulate and describe the help-seeker's feelings.	It sounds like you're feeling really frustrated and drained because your efforts don't seem to be paying off.

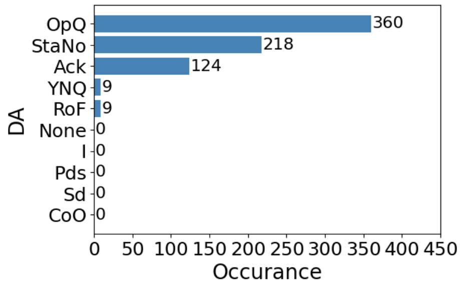
Table 4: Backward-looking strategies, definition, and example.

Strategy	Definiton	Example
YNQ	A question expecting a binary (yes/no) response.	Do you have to have any special training?
WhQ	A question beginning with a wh-word (e.g., what, who, where), seeking specific information.	Well, how old are you?
DYNQ	A statement posed as a question, expecting a yes/no answer.	So you can afford to get a house?
OpQ	A broad question inviting a wide range of responses, often conversational.	How about you?
OrC	A question offering explicit alternatives, often in the form of "or."	or is it more of a company?
CoO	A socially standard utterance used to initiate a conversation.	How are you?
Sd	Divulge similar experiences that you have had or emotions that you share with the help-seeker to express your empathy.	I completely understand how you feel. I remember feeling the same way before my first big presentation at work. I was so anxious, but I found that practicing a few extra times really helped calm my nerves.
PS	Provide suggestions about how to change, but be careful to not overstep and tell them what to do.	You can keep a note to stop your idea from going.
I	Provide useful information to the help-seeker, for example with data, facts, opinions, resources, or by answering questions.	Taking silver line from Washington D.C. to Dulles Intel Airport costs about 1 hour.

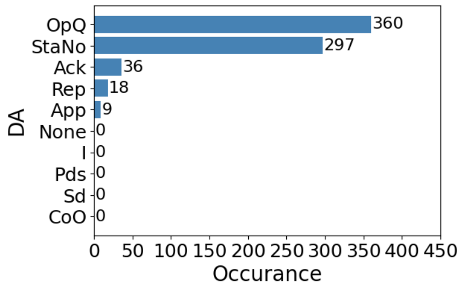
Table 5: Forward-looking strategies, definition, and example.

A.8 Personalized Dialogue Analysis

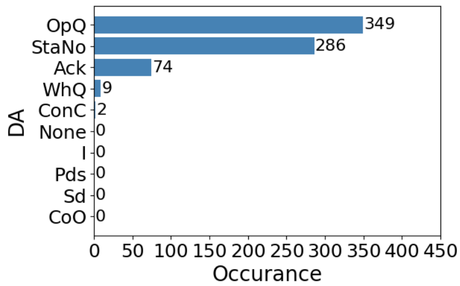
This section shows the personalized dialogue analysis. The Strategy occurrence across digital twins is shown in Figure 14. The Occurrence of each emotion transition triplet across digital twins is shown in Figure 15. The *Open Question*, *Statement-non-opinion*, and *Acknowledgment* strategies still dominate ChatWise driven conversations, suggesting their potential effectiveness in fostering engagement in conversations. The user's emotion is detected as unchanged in most triplets, indicating the difficulty of altering or measuring the user's emotional movement within a short turn.



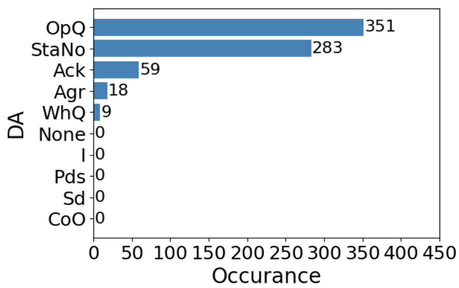
(a) Digital Twin 3.



(b) Digital Twin 5.

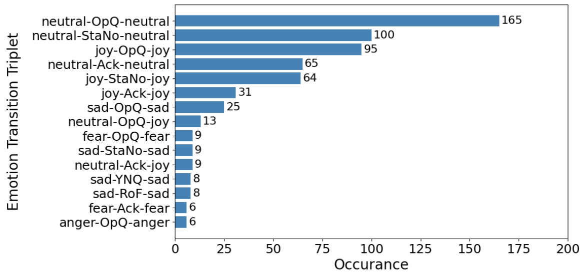


(c) Digital Twin 6.

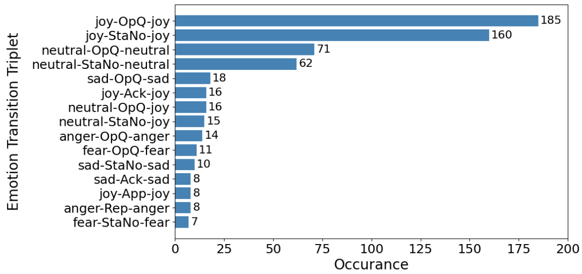


(d) Digital Twin 9.

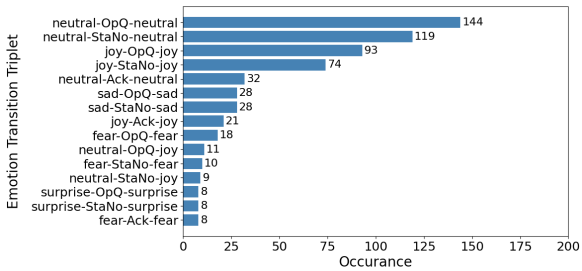
Figure 14: Strategy occurrence across digital twins.



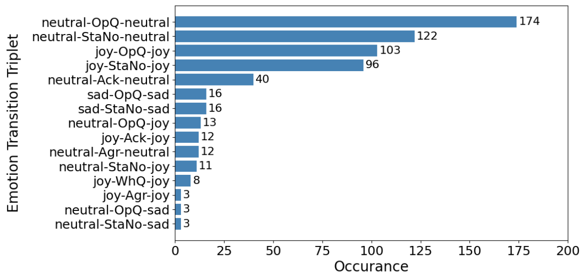
(a) Digital Twin 3.



(b) Digital Twin 5.



(c) Digital Twin 6.



(d) Digital Twin 9.

Figure 15: Occurrence of each emotion transition triplet across digital twins.