CoEmpaTeam: Enhancing Cognitive Empathy using LLM-based Avatars and Dynamic Role Play in Virtual Reality

Dehui Kong
Human-Centered Systems Lab (h-lab)
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
dehui.kong@kit.edu

Martin Feick
Human-Centered Systems Lab (h-lab)
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
martin.feick@kit.edu

Shi Liu
Human-Centered Systems Lab (h-lab)
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
shi.liu@kit.edu

Alexander Maedche
Human-Centered Systems Lab (h-lab)
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
alexander.maedche@kit.edu

Figure 1: *CoEmpaTeam* uses LLM-driven avatars with distinct personalities to foster cognitive empathy through role-switching. In Study 1, avatar self-assessment and human evaluation validated that Alice, Benji, and Caden reliably expressed their intended personalities (left). In Study 2, participants trained with *CoEmpaTeam* across three sessions over two weeks, engaging in co-living task with different avatars (center). Results showed improvements in cognitive empathy and reported transfer of these skills into everyday life (right).

Abstract

Cognitive empathy, the ability to understand others' perspectives, is essential for effective communication, reducing biases, and constructive negotiation. However, this skill is declining in a performance-driven society, which prioritizes efficiency over perspective-taking. Here, the training of cognitive empathy is challenging because it is a subtle, hard-to-perceive soft skill. To address this, we developed *CoEmpaTeam*, a VR-based system that enables users to train their cognitive empathy by using LLM-driven avatars with different personalities. Through dynamic role play, users actively engage in perspective-taking, experiencing situations through another person's eyes. *CoEmpaTeam* deploys three avatars who significantly differ in their personality, validated by a technical evaluation and an online experiment (n=90). Next, we evaluated the system through a lab experiment with 32 participants who performed three sessions across two weeks, followed by a one-week diary study. Our results showed a significant increase in cognitive empathy, which, according to participants, transferred into their real lives.

CCS Concepts

• **Human-centered computing** → **Virtual reality**;

Keywords

virtual reality, cognitive empathy, large language models, role play

ACM Reference Format:
Dehui Kong, Martin Feick, Shi Liu, and Alexander Maedche. 2026. CoEmpaTeam: Enhancing Cognitive Empathy using LLM-based Avatars and Dynamic Role Play in Virtual Reality. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3772318.3790389

1 Introduction

Cognitive empathy, often described as "knowing what another person is knowing," highlights the importance of perspective taking [33]. It plays a crucial role in fostering effective communication, reducing biases, and enabling constructive negotiation outcomes [31]. Higher levels of empathy have also been associated

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3790389

with improved quality of life and well-being across diverse contexts [9]. However, research has shown that cognitive empathy among college students has declined by nearly 40% over the past few decades [58]. While this decline is concerning, cognitive empathy is not a fixed trait but a malleable [33], learnable [89], trainable [65], and transferable [106] skill across different contexts. This combination of societal need and trainability underscores the importance of exploring new and more effective approaches to cultivating cognitive empathy.

Traditional approaches, such as classroom training programs [65, 91] and online tutorials [63, 98], have demonstrated only limited effectiveness, in part because they often lack relevance to learners' real-life social contexts. Moreover, they face challenges of scalability: classroom training requires significant resources and logistical efforts (e.g., traveling to a training site), while online tutorials are typically video- or text-based, relying on passive learning and offering little embodied experience [66]. To address these limitations, emerging immersive technologies such as virtual reality (VR) offer unique affordances, including realistic immersion [45], customizable simulation [100], and real-time interactivity [93]. These features allow users to experience another person's perspective through role play. Compared with conventional in-person role play, VR enables controlled manipulation of perspectives [55], ensuring consistent scenario delivery across participants [45] and scalable deployment without relying on multiple actors or participants [86], which enables more situated engaging forms of cognitive empathy training [4, 92, 106].

In this work, we adopt Davis's definition of cognitive empathy: *"the capacity to adopt another person's perspective and understand their thoughts and feelings without necessarily sharing their emotional state"* [23]. This perspective-taking view has roots in Smith's [102] distinction between intellectual and emotional reactions to others' experiences. While conventional role play supports perspective taking, it fixes participants into a single role, limiting opportunities to explore multiple viewpoints. To address this gap, we incorporate role-switching, a design mechanism that prompts participants to alternate between different roles, encouraging deeper reflection and a more nuanced understanding of others' perspectives.

Cognitive empathy also emerges through rich social interaction rather than in isolation [33]. Effective empathy training, therefore, requires interactive partners that can respond dynamically and contextually to human behavior. The rapid advancement of large language models (LLMs) has made it possible to convincingly simulate diverse human behaviors and generate coherent, role-consistent responses [70, 109]. LLM-driven virtual avatars can provide natural language interaction and enhance social presence [16] through multimodal expressions such as gestures, facial expressions, and tone of voice [10], offering an adaptive and immersive medium for cognitive empathy–focused interactions. To truly unlock VR's potential as an *"empathy machine,"* such capabilities should be combined with targeted perspective-taking tasks and story prompts that actively engage participants in reflection and mentalizing [64].

Given these gaps and opportunities, we frame our study around one overarching research question:

RQ: How can we design a VR system that effectively supports the development of cognitive empathy?

To address this overarching question, we examine three subquestions: (1) **RQ1**: how LLM-driven virtual avatars can be designed to exhibit distinct and coherent personalities; (2) **RQ2a**: how role-switching with LLM-driven avatars fosters cognitive empathy during VR training; and (3) **RQ2b**: how cognitive empathy skills developed during VR training translate into real-life social contexts.

To investigate these questions, we designed *CoEmpaTeam*, a novel VR application that integrates role-switching with LLM-driven avatar interaction to create structured, multi-perspective social encounters and enhance cognitive empathy. *CoEmpaTeam* immerses users in a co-living scenario where they take on one of the three roles and, together with two avatars, engage in a retrospective meeting to revisit and renegotiate house rules. The task design was inspired by the Murder Mystery Roleplay, a form of live-action role play (LARP) [49], and the avatar design was based on the Big Five personality framework [80] to capture a broad spectrum of everyday traits, with personalities validated through both LLM self-assessments and ratings from 90 participants in an online experiment. Using a mixed-methods approach and incorporating role-switching, we conducted three training sessions over the course of two weeks with 32 participants, followed by a one-week diary study. Our findings suggest that *CoEmpaTeam* not only enhanced participants' cognitive empathy during training but also facilitated the transfer of these skills into everyday life. In this work, we make four contributions:

- We present *CoEmpaTeam*, a VR-based system for training cognitive empathy, which enables users to engage in a co-living task with two LLM-driven avatars through a role-switching mechanism.
- We design three avatar roles grounded in the Big Five personality theory and validate their personality consistency through both LLM-generated self-evaluations and assessments from 90 participants.
- We conduct a three-week study with 32 participants, including three training sessions over two weeks, followed by a one-week diary study, providing empirical evidence of the effectiveness and sustained impact of *CoEmpaTeam*.
- We release *CoEmpaTeam* as an open-source system to support reproducibility, facilitate comparative evaluations, and enable extensions by the HCI research community.

2 Related Work

2.1 Approaches to Cultivating Cognitive Empathy in HCI

Cognitive empathy concerns the ability to understand and recognize others' emotions, circumstances, and experiences [6, 21]. In HCI, two approaches to fostering empathy are storytelling and role play [92]. Storytelling is a method that "immerses the audience in the experiences of characters," shaping memory and identity through shared narratives and thereby fostering empathetic understanding [79]. As a tool, storytelling has been applied across domains to enhance communication [43], build consensus and reconciliation [15], and foster empathetic public engagement [74].

Beyond narrative immersion, role play provides an embodied pathway to empathy. By enacting different roles, participants can experience diverse perspectives [34] and enhance perspective-taking

[19]. Research has shown that empathy is more likely to emerge when participants enact roles dissimilar to themselves [81]. However, the specific demands and situational constraints of different roles may also lead to divergent perspectives. Prior work suggests that exchanging social positions (i.e., role-switching) can help reconcile these divergences and deepen understanding [35]. Moreover, role play generates episodic memories of lived experiences, which, when contrasted through role-switching, can further foster cognitive empathy [105]. However, traditional role-play approaches typically center on a single role, restricting systematic perspective-switching across roles and limiting sustained opportunities for cognitive empathy training.

2.2 Role of VR in Cognitive Empathy Training

VR has been increasingly applied in HCI, demonstrating unique advantages in challenging contexts such as mental health [70], social inclusion [28], and cross-cultural understanding [92]. In recent years, VR has also been incorporated into empathy cultivation across diverse domains, including interfaith learning [92], crisis and illness experiences [60, 116], human–animal relations [114], cultural heritage [118], and intergenerational communication [99]. Despite the diversity of contexts, these applications emphasize immersive role play, contextualized interaction, and narrative guidance as core design elements to evoke cognitive engagement [82].

VR has often been described as an *"empathy machine"* [40, 42], yet its potential for cultivating empathy arises not from any single capability but from the integration of multiple affordances. First, realistic immersion enables users to inhabit specific perspectives and contexts [45], situating understanding within concrete social and cultural environments [92]. Second, customized simulation allows designers to guide attention through spatial arrangements and narrative cues, helping participants engage with unfamiliar experiences and "step into others' shoes" [100]. Finally, real-time interactivity provides a safe environment for practice and exploration [93]. Compared to abstract representations, such embodied and concrete experiences are more likely to elicit critical reflection and empathic understanding, which can be deepened through repeated interaction [57].

Beyond these affordances, empirical studies provide evidence that VR role-switching can translate empathic experiences into real-world attitudes and behaviors [75, 106]. For instance, role exchange in school bullying scenarios has been shown to foster moral reasoning and willingness to help [38], while perspective shifts in police training enable trainees to move from the "perpetrator" role to the "victim" role, supporting experiential learning and empathy cultivation [55]. These studies indicate that cognitive empathy is not only influenced by environmental context but can be intentionally cultivated through targeted task design. To fully realize VR's potential, design must therefore go beyond immersion and narrative fidelity to integrate explicit cognitive interventions that scaffold perspective-taking and critical reflection [64]. In this view, effective empathy training in VR is not about "watching a story" but about "inhabiting a role, driving the process, and making decisions."

2.3 LLM–Driven Social Avatars in VR

Recent advances in LLMs have enabled their integration into VR, significantly enhancing interactive experiences in digital environments [1, 84]. Researchers have applied LLM-driven VR systems across domains such as education [14, 72], accessibility [20, 67], and training [27, 71]. For instance, *ClassMeta* introduced adaptive classroom avatars that provide personalized learning support [72], while *GlanceWriter* leveraged gaze-driven avatars to assist users with motor impairments in writing tasks [20]. In training contexts, generative social simulations have been used to create realistic social scenarios that support self-care practice under stress [27]. Across many of these domains, LLM-driven avatars have emerged as the central medium for interaction, capable of generating context-relevant dialogues by integrating system prompts, conversational memory, and user input [73]. These developments build on a longer trajectory of virtual agent research. Early systems evolved from scripted characters [51] to emotionally driven agents such as *FearNot!*, which generated emergent narratives to foster empathy in anti-bullying education [2]. More recently, generative approaches such as *Generative Agents* have demonstrated how LLM-powered characters can sustain daily routines and interpersonal relationships in sandbox environments [87]. This trajectory from early scripted logic to contemporary LLM-powered avatars reflects a broader shift toward open-ended, generative modes of interaction, while recent work has also begun extending these capabilities into more structured training scenarios.

Building on this trajectory, LLM-driven avatars can dynamically adjust narratives, emotions, and motivations, thereby delivering more immersive and authentic interactive experiences [70, 119]. In social VR, they enhance presence and social realism [39, 101]; in medical and training contexts, they foster engagement through emotionally rich feedback [119]. Beyond language generation, their credibility also depends on multimodal cue integration, visual appearance, and personality design [13, 107]. Building on these design considerations, recent studies have explored applications ranging from virtual patient simulators [119] to virtual classroom companions [72] and language-learning assistants [85], demonstrating the potential of LLM-driven avatars across diverse contexts.

In terms of interaction forms, LLM-driven avatars have expanded from text to multimodal expressions such as gestures, facial expressions, and gaze, enabling more immersive and credible interactions [83, 90]. This shift is pushing VR avatars beyond scripted dialogues toward dynamic, context-aware, and emotionally resonant modes of interaction. A central mechanism is the temporal and spatial alignment of verbal and visual cues. Such alignment supports conversational fluency and spatial orientation, reduces cognitive load through sensory consistency, and enhances task performance [17, 77]. In VR environments, these multimodal interactions not only improve intuitiveness but also foster stronger emotional and social awareness, thereby deepening user engagement [39, 40]. However, despite these advances, existing LLM-driven avatar systems rarely adopt distinct personalities across multiple avatars, limiting their capacity to present differentiated perspectives.

Overall, prior work highlights the potential of role play, VR, and LLM-driven avatars for empathy training, yet these elements have rarely been integrated into a unified and systematic framework.

Building on these insights, we introduce *CoEmpaTeam*, a VR system that unites multi-avatar interaction with dynamic role-switching to advance situated training for cognitive empathy.

3 Developing *CoEmpaTeam*

CoEmpaTeam is an immersive VR application in which participants engage in role play with two LLM-driven avatars, each embodying distinct personalities within a designed living situation. The scenarios situate participants in everyday, conflict-prone contexts that require communication, negotiation, perspective-taking, and joint decision-making. By switching roles among these characters, participants directly experience multiple perspectives and practice reconciling differences through dialogue and negotiation, which are essential for understanding, anticipating, and responding to others' perspectives in real-world social contexts. Through this process, *CoEmpaTeam* aims to cultivate users' cognitive empathy.

3.1 Design

3.1.1 Co-living Task. We designed *CoEmpaTeam* to immerse participants in a familiar yet conflict-prone context—shared living arrangements [18, 29]. This choice is motivated by the fact that co-living often involves disagreements over noise, cleanliness, kitchen use, guest policies, and personal boundaries [18, 29, 54]. Such everyday conflicts are highly relatable to daily experiences while also laden with tensions around value differences and responsibility allocation. Inspired by role-playing games such as Murder Mystery Roleplay [5, 49], we translated this scenario into a co-living task, where participants and two avatars hold a retrospective house meeting to decide whether to continue co-living and how to renegotiate house rules (see Supplement).

In this process, participants must negotiate from multiple standpoints, actively engaging in perspective-taking. Psychological research regards perspective-taking as a critical mechanism for cultivating cognitive empathy [23], as it enhances self–other overlap and facilitates social coordination [31]. Building on this foundation, we hypothesize that the everyday conflicts embedded in *CoEmpaTeam* provide an effective basis for eliciting perspective-taking, thereby fostering the development of cognitive empathy.

3.1.2 Role Play. *CoEmpaTeam* extends conventional role play with a role-switching mechanism that requires participants to sequentially enact three different characters across multiple sessions. Before each session, participants receive detailed role cards containing Basic Info, Lifestyle Log, Hidden Motivation, and Stance on House Rules (see Supplement), which provide sufficient context for role immersion. By alternating between roles, participants move beyond a single viewpoint and are encouraged to confront conflicting perspectives. This design transforms role play into a structured multi-perspective exercise, prompting reflection, re-examination of assumptions, and deeper understanding of others' positions, thus operationalizing prior findings on perspective-taking as a concrete mechanism for cultivating cognitive empathy [31].

3.1.3 Avatar Design. Personality as a driven variable was central to the design of the avatars in the *CoEmpaTeam*. The Big Five personality framework [36], widely validated in psychology and increasingly adopted in HCI to capture individual differences [11, 50], served as the basis for our design. Each avatar was modeled with a distinct personality profile based on the Big Five [36], varying in openness, conscientiousness, extraversion, agreeableness, and neuroticism (see Supplement). This variation was designed to foster diverse perspectives, communication styles, and interpersonal tensions, conditions that are critical for eliciting cognitive empathy and perspective-taking, which form the training goals of the task. To operationalize these traits, we employed LLM prompting strategies. Following a zero-shot learning approach [10], we crafted instructions that implicitly conveyed the desired traits, guiding the model to generate responses aligned with specific personality characteristics. Prior work demonstrates that descriptive prompts can elicit stable expressions of personality by framing the interaction context in ways that encourage the model to naturally embody the intended persona [48]. Our implementation builds on these insights to ensure that avatar dialogue consistently reflects distinguishable yet coherent personalities.

To situate these personalities within a manageable yet socially rich context, the avatars were arranged in a three-person group. Triads represent one of the fundamental "core configurations" of human social interaction [12] and strike an effective balance between complexity and cognitive load [103], providing enough diversity for perspective-taking while avoiding cognitive overload. For narrative coherence, avatars were assigned fixed genders (two females, one male). While gender was not treated as a study variable, it remains unclear whether this choice influenced participants' perceptions. Future work should examine how different gender configurations may shape experiences in similar settings.

To further enhance immersion, we integrated verbal and nonverbal cues into avatar behavior. Prior work highlights that the combination of speech and embodied signals is essential for creating natural and engaging human–agent interactions [13]. Building on this, the LLM output was structured into four fields: speaker, text, emotion, and gesture. These were mapped to avatar behaviors (see 3.1.4), ensuring that linguistic content was consistently accompanied by expressive cues and that verbal output aligned with corresponding emotional and gestural expressions. To avoid avatars appearing static or mechanical, we further introduced micro-behaviors such as random blinking and subtle eye movements, thereby enhancing social presence and strengthening both clarity and expressiveness in interaction [113].

Figure 2: Avatars used in the *CoEmpaTeam* system. From left to right: Alice, Benji, and Caden.

3.1.4 Implementation. *CoEmpaTeam* was implemented in Unity3D (2022.3.51f1)[1] and deployed on the Meta Quest Pro headset[2]. The

[1] https://unity.com/
[2] https://www.meta.com/quest/quest-pro/

virtual environment was set in a kitchen to reflect the domestic context of co-living. Most 3D assets were sourced online and optimized to balance authenticity with the limited computational resources of standalone VR devices. Avatar models were generated using Ready Player Me[3] (see Figure 2).

During interaction, participants' speech was transcribed into text using a locally deployed OpenAI Whisper model[4]. Transcriptions were appended to the dialogue history and forwarded to a Python backend, where structured prompts were constructed. These prompts incorporated task background, turn-taking rules, output format, and avatar-specific personas, including their possible gestures and emotional repertoires (see Supplement). The backend then queried the Llama 3.1 8B Instruct model[5] to generate a response. The structured output was sent to the Unity client, where it was parsed to drive avatar behaviors. Each response was structured into four fields: *speaker*, *text*, *gesture*, and *emotion*. The *text* field was converted into natural speech via ElevenLabs TTS[6], with articulation synchronized using the Oculus LipSync Unity SDK[7]. *Gestures* were animated with Mixamo[8], while *emotions* were mapped to avatar-specific facial blendshapes to reflect distinct personalities. To further enhance social presence, avatars were equipped with dynamic gaze control using Unity Final IK (Look At IK)[9], enabling them to orient toward the current speaker. Code and the Unity package are available online[10].

3.1.5 CoEmpaTeam Role-play Workflow. Upon entering the environment, participants selected a role (see Figure 3a) and reviewed information about their assigned character, including Big Five personality traits, lifestyle logs, and hidden motivations (see Figure 3b). To support role-play, they were also given basic information about the two other avatars (see Figure 3d). Embodiment was reinforced through motion-synchronized avatar arms (see Figure 3e-f), enhancing visuomotor alignment and presence [104]. This design was informed by the Proteus effect, which suggests that embodying a virtual character can shape users' attitudes and behaviors [115].

To initiate the task, participants clicked "start." After each speech turn, they pressed "finish speaking" to transmit their input to the system (see Figure 3b), which in turn triggered responses from the two avatars. This interaction loop structured the role-play while maintaining flexibility, ensuring consistent turn-taking and supporting naturalistic conversations with the avatars.

3.2 Study 1: Validation of Avatars

To address RQ1, we combined avatar self-assessment with human evaluation to validate whether the three designed avatars exhibited behaviors consistent with their intended personalities.

3.2.1 Avatar Self-assessment. Following [61], we administered the NEO-FFI-30 personality inventory [59] to each avatar, repeating

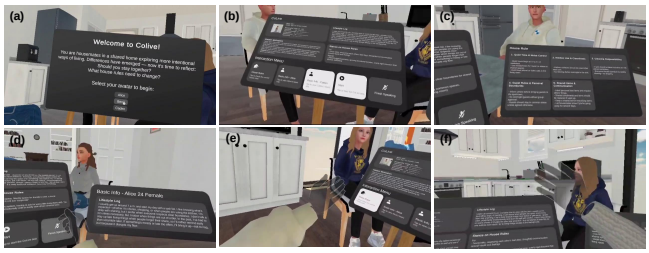


Figure 3: *CoEmpaTeam* system interface and role-play workflow: (a) Welcome screen for avatar selection. (b) Role-specific information with options to start the task or finish speaking. (c) Shared house rules organized into five categories. (d) Example of basic information for Alice to support role-play. (e-f) Motion-synchronized avatar arms when embodying Benji and Caden (Alice analogous).

Table 1: Big Five trait profiles of avatars across 100 NEO-FFI-30 trials ($M \pm SD$), with cosine similarity to target profiles.

	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism	Cosine Similarity
Alice (Target)	13	23	13	15	6	–
Alice ($M \pm SD$)	13.83 ± 1.38	22.95 ± 0.28	14.83 ± 0.64	14.39 ± 1.22	6.16 ± 0.83	0.997 ± 0.002
Benji (Target)	20	9	17	13	12	–
Benji ($M \pm SD$)	19.79 ± 1.12	8.94 ± 1.17	16.81 ± 1.14	13.04 ± 1.09	12.03 ± 1.06	0.862 ± 0.017
Caden (Target)	14	17	8	20	10	–
Caden ($M \pm SD$)	13.79 ± 1.12	16.88 ± 0.67	7.81 ± 1.14	20.04 ± 0.67	10.03 ± 1.06	0.950 ± 0.010

the procedure 100 times to account for variability in model outputs. We calculated mean scores (M) and standard deviations (SD) for the Big Five traits, as well as cosine similarity to the intended profiles as a measure of overall fidelity (as detailed in Table 1). Alice's responses were highly consistent with her predefined profile (cosine similarity $M = 0.997$, $SD = 0.002$), reflecting her structured and rule-oriented persona. Benji's profiles, in contrast, showed greater variability ($M = 0.862$, $SD = 0.017$), aligning with his more flexible and less organized character. Caden fell between these extremes, exhibiting high fidelity ($M = 0.950$, $SD = 0.010$) while maintaining slight natural variation. Together, these results demonstrate that the avatars not only reproduced their intended Big Five configurations but also expressed variability patterns consistent with their designed personalities.

3.2.2 Human Evaluation. To evaluate whether the avatars' overall behaviors aligned with their intended personalities, we conducted a between-subjects online study. 90 participants were randomly assigned to one of three conditions (focusing on either Alice, Benji, or Caden), with 30 participants evaluating each avatar. Each participant watched a five-minute pre-recorded video of a discussion round between the avatars before completing a third-person version of the NEO-FFI-30 personality inventory [59] and an additional five-item Likert questionnaire (1–5) assessing the perceived consistency between the avatar's verbal expressions, non-verbal behaviors, and vocal style (see Supplement).

We recruited 90 participants (42 male, 48 female) through Prolific[11], an online participant recruitment platform. Each participant received €3 as compensation. To avoid potential transfer effects

[3]https://hub.readyplayer.me/avatar/choose
[4]https://github.com/openai/whisper
[5]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[6]https://elevenlabs.io/app/developers/api-keys
[7]https://developers.meta.com/horizon/documentation/unity/audio-ovrlipsync-unity/
[8]https://www.mixamo.com/
[9]https://docs.readyplayer.me/ready-player-me/integration-guides/unity/setup-for-xr-beta/setup-final-ik
[10]https://github.com/kindhui62/CoEmpaTeam
[11]https://www.prolific.com

from prior exposure to the avatars, none of the participants in this validation study took part in the subsequent main experiment. On average, the study took 12 minutes to complete.

To ensure data quality and align with established gold-standards for crowdsourced research [30, 56], the questionnaire included a attention-check item (e.g., "Please select 'strongly agree' for this statement"), which all retained submissions passed. We also examined completion times to identify potential low-effort responses, such as unusually short durations; no such anomalies were detected.

Participants first watched a five-minute pre-recorded video in which the three avatars discussed the co-living task. In each video, the target avatar was placed in the center of the screen, with the other two positioned to the sides to simulate a realistic group interaction while directing participants' attention to the evaluated character (see Figure 4). After viewing, participants completed the questionnaires described above.

Figure 4: Example frames from the evaluation videos. Each condition featured a different target avatar placed in the center of the scene: (a) Alice, (b) Benji, and (c) Caden. The other two avatars were positioned to the sides to simulate a realistic group interaction while directing participants' attention to the evaluated role.

3.2.3 Results. We collected 90 complete questionnaire responses. For each avatar, we aggregated personality ratings and compared them with the predefined trait profiles. To evaluate reliability, we assessed the third-person version of the NEO-FFI-30 [59]. Across avatars, the subscales showed acceptable to good internal consistency (Cronbach's α = .67–.87), with most exceeding the conventional .70 threshold. The only exception was Openness for Alice (α = .67), which remains acceptable for short scales such as the NEO-FFI-30 [97].

We then analyzed the alignment between perceived and predefined personalities using two complementary measures: (1) Pearson's r, capturing whether the relative ranking of trait scores across the Big Five was consistent, and (2) Cosine similarity, assessing the overall configurational similarity between the perceived and predefined profiles as vectors. Pearson's r thus captures trait ranking alignment, while cosine similarity reflects overall configurational fidelity. Applying these measures, we found that participants' perception of Alice was strongly and significantly correlated with the intended profile (r = .93, p < .05). Benji's ratings also indicated a strong correlation (r = .87, p = .052). Similarly, Caden's perceived profile demonstrated a strong and significant correlation (r = .89, p < .05). Cosine similarity analyses confirmed the overall alignment: Alice (M = 0.969, SD = 0.027), Benji (M = 0.959, SD = 0.025), and Caden (M = 0.967, SD = 0.036). Finally, we calculated mean (M) and standard deviation (SD) scores for each item of the verbal–nonverbal and vocal consistency questionnaire. Ratings were positive overall, with mean scores ranging between 3.5 and 4.1 (out of 5) (see Supplement). Participants rated the avatars highest on items concerning the match between voice and personality as well as the overall believability of the character (Q3–Q4). Overall, these findings indicate that participants perceived all three avatars as internally consistent and believable, providing validation of the character designs.

3.2.4 Summary. Our dual evaluation, combined with the avatar self-assessment and human evaluation, confirms that the avatars consistently expressed their intended personalities, Alice as structured and rule-oriented, Benji as flexible and less organized, and Caden as accommodating and harmony-seeking. This dual validation establishes the credibility of the avatar designs, ensuring the avatars are both computationally coherent and experientially believable. This provides a solid foundation for the subsequent cognitive empathy training study.

4 Study 2: Cognitive Empathy Training

To address RQ2a and RQ2b, we conducted a three-week study in which participants completed three role-switching sessions across two weeks (RQ2a), followed by a one-week diary study (RQ2b), to examine how repeated role-switching with *CoEmpaTeam* fosters cognitive empathy and how these skills transfer into everyday life.

4.1 Study Design

Each participant engaged in three training sessions, role-playing Alice, Benji, and Caden. To account for the distinct traits of the roles while minimizing potential order effects, we employed a 3×3 Latin square design that counterbalanced the sequence in which participants enacted the roles. Following prior VR empathy research in HCI [60, 92], which evaluates systems holistically, we likewise examine *CoEmpaTeam* as an integrated training experience rather than isolating individual components.

The training phase lasted two weeks, with sessions scheduled at least two days apart to encourage reflection while maintaining continuity. At the beginning of each session, participants were screened for VR-related motion sickness and mental health issues to ensure safe participation. Completion of all three sessions was required for inclusion in the study. Before and after the training sessions, participants completed questionnaires to assess cognitive empathy and related constructs (see 4.4). After the last training session, we invited a subset of participants for a 20-minute semi-structured interview about their experiences with *CoEmpaTeam*.

To measure the real-world transfer of cognitive empathy training, participants completed a three-entry diary study (one entry every two days) during the week after the training. Diary methods are particularly valuable in HCI research for their ecological validity, as they enable in-situ reflection on real-world experiences beyond the controlled lab environment [22].

4.2 Participants

We recruited 32 participants (22 male, 10 female) from a local university, aged 18–34 (M = 24.00, SD = 3.41). All attended the first session, 27 returned for the second, and 23 completed the third. The final analysis included 22 participants (14 male, 8 female, M = 24.39, SD = 3.71) who completed all three training sessions. One additional participant was excluded due to exceeding the cut-off on the

General Health Questionnaire–12 (GHQ-12) [25], consistent with our pre-defined screening criteria. Based on participants' feedback, attrition appeared to be mainly due to scheduling conflicts (e.g., overlapping classes or exams), rather than study-related discomfort, though other contributing factors cannot be fully ruled out. The cultural backgrounds of participants were diverse, with 9 participants from Western Europe, 3 from Eastern Europe, 4 from East Asia, 5 from South Asia, and 1 from the Middle East. Regarding prior VR experience, 2 participants had never used VR, 14 had used it once or twice, and 6 reported occasional use a few times per year. None reported frequent VR use.

Participants received increasing compensation across sessions (€13, €15, €17) to acknowledge their time commitment, with additional payments for the interview (€5) and for each of the three diary entries (€3, €4, €5). The study was approved by the institutional ethics and data protection committee, and all procedures adhered to applicable data protection regulations. Participants were informed that participation was voluntary across all stages of the study and that they could withdraw at any time without penalty.

4.3 Procedure

4.3.1 Training Sessions. At the beginning of the study, each participant was assigned to an individual room to ensure a quiet environment. After providing informed consent, participants completed a 15-minute onboarding session to familiarize themselves with the system. They then received printed role materials describing the co-living task, including background information, task instructions, house rules and a role card outlining personality traits, lifestyle log, hidden motivations, and stance on house rules (see Supplement). To support role immersion, we also provided an overview of the Big Five personality dimensions (see Supplement). Participants subsequently engaged in a short practice phase to get familiar with the VR headset and system.

Each participant completed three training sessions across two weeks. At the beginning and end of each session, participants filled out pre- and post-test questionnaires (see 4.4.1, 4.4.2). During the session, they wore a Meta Quest Pro headset, selected their assigned role, and engaged in a 20-minute discussion on household rules with two other LLM-driven avatars, each with distinct personalities. The order of roles (Alice, Benji, and Caden) was counterbalanced across participants.

4.3.2 Post-Training Interview & Diary. After completing all three sessions, participants were invited to a voluntary 20-minute semi-structured interview focusing on their overall experience with the *CoEmpaTeam* system, engagement in cognitive empathy-related practices, and perceptions of the role-play and avatars. In the following week, they participated in a diary study consisting of three online entries (one every two days), guided by open-ended prompts. The diaries encouraged participants to reflect on empathy-related experiences in their daily life and to consider possible links to the VR sessions (see 4.4.3).

4.4 Data Collection

4.4.1 Pre-test Questionnaire. At the beginning of the study, participants provided demographic information (age, gender, nationality, and prior VR experience). Before each training session, we administered the General Health Questionnaire (GHQ-12) [25] to screen for mental health risks and the Simulator Sickness Questionnaire (SSQ) [53] to monitor VR-related motion sickness. These measures were used solely for screening and did not enter the analysis.

To assess cognitive empathy, we used the Interpersonal Reactivity Index (IRI) [24], a widely used self-report measure of empathy. The IRI consists of four subscales: Perspective Taking, Empathic Concern, Fantasy, and Personal Distress. Consistent with our research focus, we examined the two subscales most directly linked to cognitive empathy: Perspective Taking (PT) and Fantasy (FS). PT captures the tendency to adopt others' viewpoints in everyday life, whereas FS reflects the tendency to imaginatively identify with fictional characters, paralleling the role-taking required in our VR training. Both subscales have demonstrated high internal consistency in prior research (PT: Cronbach's $\alpha = .83$; FS: $\alpha = .86$) [52].

Participants responded to 14 items on a 5-point Likert scale (1 = does not describe me well, 5 = describes me very well), with higher scores indicating stronger cognitive empathy. Example items include *"I try to look at everybody's side of a disagreement before I make a decision"* (PT) and *"I often find myself imagining what it would be like to be in the shoes of a character in a book/movie."* (FS). The IRI has been widely used to detect changes in empathy using paired t-tests [23, 112] and has also been applied in VR research to investigate cognitive empathy and perspective-taking [45, 106], making it a well-established and reliable measure for our study.

4.4.2 Post-test Questionnaire. After each training session, participants completed several measures. To assess changes in cognitive empathy, we administered the IRI [24] again. To evaluate experiences in the virtual environment, we included the Presence Questionnaire (PQ, 5-point Likert scale) [111] and the ownership subscale of the Virtual Embodiment Questionnaire (VEQ, 5-point Likert scale) [37]. To explore whether the avatars elicited an uncanny valley effect, we added the Uncanny Valley Index (UVI, 7-point Likert scale) [46]. Finally, after the third session, participants completed the NASA Task Load Index (NASA-TLX, 7-point Likert scales) [41] to examine potential task-related strain.

4.4.3 Qualitative Data. To complement the quantitative measures, we collected qualitative data through semi-structured interviews and diary entries to capture participants' perspectives in depth. The post-training interviews explored participants' overall experience with the system, their engagement in cognitive empathy (e.g., perspective-taking and role-switching), their sense of role immersion, perceptions of the avatars, and reflections on possible improvements to the empathy training mechanisms. In the following week, participants completed a diary study consisting of three entries (one every two days). Each entry included four open-ended prompts and followed a shared structure: participants were asked to recall a recent real-world interpersonal interaction, describe how they engaged in perspective-taking, and reflect on whether any aspects of the VR training informed their experience. While the overall structure was consistent, each entry varied slightly in reflective emphasis. The first entry focused on recalling a recent interaction and describing how they attempted to understand another person's viewpoint. The second centered on how participants adapted their responses during interactions and whether role-switching

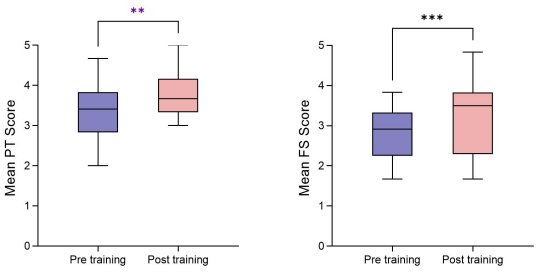
in VR influenced this process. The third invited broader reflection on changes over time, including perceived changes in cognitive empathy related behaviors and intentions to continue applying related strategies. The full set of diary prompts is included in the Supplement. To ensure data quality, participants were reminded that there were no right or wrong answers, that compensation was based solely on completion rather than content. Together, the interviews and diaries captured participants' reflections on cognitive empathy-related experiences inside and outside VR. We analyzed all qualitative data using thematic analysis [8], following an inductive coding approach [94] to identify recurring themes.

4.5 Results

We report both quantitative and qualitative findings. Quantitatively, we focus on changes in cognitive empathy using the Interpersonal Reactivity Index (IRI), alongside measures of presence (PQ), embodiment (VEQ), uncanny valley perceptions (UVI), and task load (NASA-TLX). Qualitatively, our analysis of the interviews and diary entries provides deeper insights into participants' training experiences and the transfer of empathy strategies into daily life.

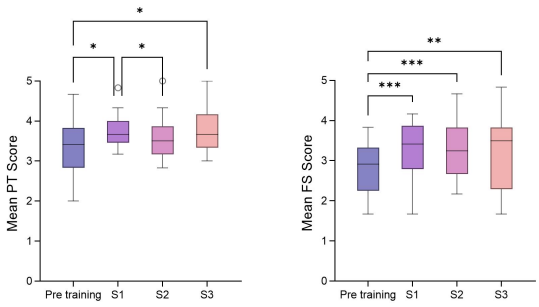
4.5.1 *Quantitative Results.*

Cognitive Empathy: Perspective Taking and Fantasy (IRI). The IRI sample included 22 participants. In line with our focus on cognitive empathy, we analyzed the Perspective Taking (PT) and Fantasy (FS) subscales separately. Both subscales demonstrated high internal consistency in our sample (Cronbach's $\alpha = .77–.86$; FS: $\alpha = .79–.87$). Paired-sample t-tests revealed significant improvements on both measures (see Figure 5). PT increased from a pre-training mean of $M = 3.36$ ($SD = 0.65$) to $M = 3.70$ ($SD = 0.51$), $t(21) = 2.98$, $p = .007$, Cohen's $d = 0.64$ (medium-to-large). FS increased from $M = 2.76$ ($SD = 0.64$) to $M = 3.22$ ($SD = 0.92$), $t(21) = 4.04$, $p < .001$, $d = 0.86$ (large). These results indicate that participants demonstrated stronger perspective-taking and imaginative role engagement after training. In psychological research on human behavior, effect sizes typically range from small to medium ($d = 0.2–0.5$) [95], our observed effects ($d = 0.64–0.86$) exceed these benchmarks, suggesting that the *CoEmpaTeam* training enhanced cognitive empathy.



(a) PT (b) FS

Figure 5: Pre- and post-training changes in IRI subscales. Both PT and FS increased significantly.



(a) PT (b) FS

Figure 6: Session-wise IRI scores. PT and FS were elevated relative to pre-training; PT showed session-to-session fluctuations, whereas FS showed no significant differences among the three training sessions.

To examine the trajectory of empathy development across sessions, we conducted repeated-measures ANOVAs on PT and FS (see Figure 6). For PT, there was a significant main effect of session, $F(2.11, 44.25) = 7.40$, $p = .0014$, partial $\eta^2 = .26$. Post-hoc comparisons showed that PT increased significantly from pre-training to Session 1, decreased significantly from Session 1 to Session 2, and showed a non-significant increase from Session 2 to Session 3. Despite these fluctuations, PT scores in all training sessions remained higher than at pre-training. FS also showed a significant main effect of session, $F(2.66, 55.84) = 9.51$, $p < .001$, partial $\eta^2 = .31$. Post-hoc comparisons indicated that FS was significantly higher in all three training sessions compared to pre-training, with no significant differences among Sessions 1, 2, and 3.

Overall, these findings indicate that cognitive empathy improved significantly from pre- to post-training. Session-level analyses further reveal a pattern of early gains in both PT and FS, followed by session-to-session fluctuations in PT, while FS remained consistently elevated across all training sessions relative to pre-training.

System Experience: Presence, Embodiment, and Avatar Perception (PQ, VEQ, UVI). To assess participants' experience in VR and their perception of the avatars, we analyzed three measures. Presence (PQ) scores remained at a moderate level across the three sessions (S1: $M = 3.06$, $SD = 0.42$; S2: $M = 3.20$, $SD = 0.56$; S3: $M = 3.21$, $SD = 0.55$), showing a slight upward trend but no significant differences ($p > .05$) (see Figure 7a). This indicates a stable and sufficient sense of "being there" in VR, maintained as participants became familiar with the system. Similarly, embodiment scores on the VEQ ownership subscale remained moderate across sessions (S1: $M = 2.46$, $SD = 0.54$; S2: $M = 2.54$, $SD = 0.49$; S3: $M = 2.56$, $SD = 0.57$), with no significant differences ($p > .05$) (see Figure 7b). This limited ownership likely reflects the restricted mapping of movements to the avatar, which included arm control and natural head tracking but not full-body animation. It may also relate to the system's emphasis on social rather than physical engagement. Nevertheless, the stable ratings suggest that participants could still relate to the avatars in socially meaningful ways. Embodiment may be enhanced in future iterations of *CoEmpaTeam* by incorporating

richer body tracking, gesture recognition, or more expressive nonverbal cues. By contrast, UVI ratings showed no evidence of the classic uncanny valley dip. Instead, comfort increased monotonically with perceived humanness (see Figure 7c). This pattern was consistent across sessions, indicating that participants adapted positively to the avatars' appearance and experienced them as socially acceptable within the training context.

Perceived Workload (NASA-TLX). The NASA-TLX comprises six subscales: Mental Demand (MD), Physical Demand (PD), Temporal Demand (TD), Performance (P), Effort (E), and Frustration (F). Participants rated each subscale on a 7-point Likert scale. For analysis, the Performance scale was reverse-coded so that higher values indicate better performance, while for the other subscales lower values indicate lower workload. Mean scores indicated moderate mental demand ($M = 3.04$, $SD = 1.63$), low physical ($M = 1.50$, $SD = 0.72$) and temporal demand ($M = 2.46$, $SD = 1.56$), high perceived performance ($M = 4.33$, $SD = 2.01$), moderate effort ($M = 3.33$, $SD = 1.43$), and low frustration ($M = 2.21$, $SD = 1.14$) (see Figure 8). Overall, participants were cognitively engaged in the task while experiencing little physical or temporal strain. They perceived their performance as relatively high and reported low levels of frustration, suggesting that the training was demanding enough to elicit involvement without imposing excessive workload.

4.5.2 *Qualitative Findings from Interviews.*

Participants completed the training individually. For logistical reasons, two participants were scheduled in the same time slot but worked in separate rooms, while interviews were conducted one-on-one by a single researcher. As a result, only one participant per slot could be interviewed immediately after the session. To obtain a balanced and representative subset, we selected interviewees across time slots while considering gender and cultural diversity. This resulted in 11 completed interviews (P01–P11; $M = 24.27$, $SD = 3.74$; 6 male, 5 female), with cultural backgrounds spanning Western Europe (4), Eastern Europe (1), East Asia (2), South Asia (3), and the Middle East (1). Each interview lasted approximately 20 minutes. The analysis revealed four recurring themes: Clear frameworks facilitated reflective and enjoyable experiences, design features enabled natural and deeper role adoption, role-switching fostered perspective expansion, and avatars and environments supported empathic engagement. We elaborate on each theme below with illustrative quotes.

Clear Frameworks Facilitated Reflective and Enjoyable Experiences. Participants consistently characterized *CoEmpaTeam* as both enjoyable and reflective. As P01 noted, *"I really enjoyed it, it was a fun experience."* They highlighted that the system's clear framework for discussion made the training feel organized rather than arbitrary, which helped them stay attentive and engaged. As P07 noted, *"[The setup] made me stop and think about what the character would do, rather than just answering as myself."* Similarly, P05 explained that the experience *"was very easy, and as the sessions progressed, I think it was easier for me to adopt [the role]."* Several participants further appreciated that the task provided a clear framework for dialogue, which they felt supported constructive and thoughtful exchanges. While we ensured that all participants received introductory training to become comfortable with the VR environment, some still felt that their limited familiarity with VR constrained the meaningfulness of the experience. To support novice users, participants suggested adding more intuitive interface cues (P07) and short introductory videos describing the roles to help users settle into the role (P05). Overall, participants valued the enjoyable and reflective nature of the training, as the clear framework for discussion helped them stay attentive while considering different perspectives. They also suggested improvements to ease onboarding and further enhance the experience in future iterations.

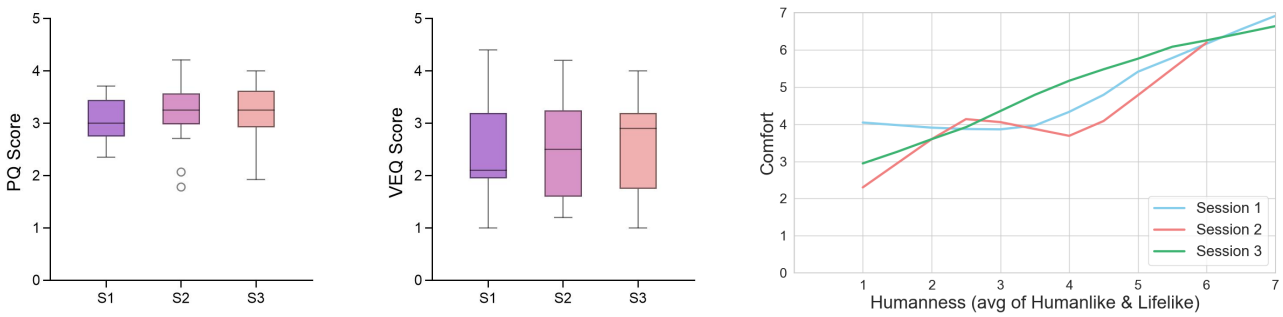
Design Features Enabled Natural and Deeper Role Adoption. Participants generally indicated that they were able to step into the assigned roles naturally. For example, P05 noted it was *"like playing a theater role."* This made role embodiment feel accessible and engaging, allowing them to move beyond their habitual ways of speaking or responding.

Role cards and hidden motives were repeatedly highlighted as essential tools for role adoption. They provided clear background information that anchored participants in the perspective of the role. As P01 explained, *"I was able to do that very naturally because the role cards were very clear."* They further noted that such traits often resembled people they had encountered in real life, which made the role easier to inhabit. Similarly, P07 added that *"it was pretty easy to follow along because we had the whole [role cards]."* Hidden motives, in particular, encouraged deeper reflection on how to enact a role. P11 noted, *"when I realized the character had a hidden reason, I thought harder about how to behave."* Collectively, these design features supported participants in moving beyond surface-level enactment toward more authentic role immersion.

Role-Switching Fostered Perspective Expansion. Participants experienced role-switching as both engaging and enlightening. Some likened the activity to a game that nonetheless prompted them to consider others' perspectives (P01), while others compared it to writing a novel that required imagining characters' motivations beyond their own (P11). These metaphors illustrate how role-switching combined playful engagement with deliberate perspective-taking, combining enjoyment with cognitive work.

Role-switching also heightened sensitivity to contexts and interpersonal differences. P04 explained that inhabiting different roles forced him to *"think about what others would do in this situation,"* which cultivated greater contextual awareness. Similarly, P09 reflected that *"people think in completely different ways than I do, and I need to pay attention to that,"* noting that fictional roles even reminded him of real-life acquaintances. Such realizations prompted participants to adapt their own preferences or reconsider habitual responses. For instance, P01 recalled adjusting his choices as Benji to accommodate Alice's early waking habits. Embodying contrasting perspectives side by side highlighted the diversity of thought and behavior, pushing participants to practice perspective-taking beyond what they typically attempted in daily life.

These insights were reinforced through repeated exposure across sessions. P04 noted that sequentially playing multiple roles helped him better grasp the dynamics between roles: *"By switching the roles, I could understand each of them more, when I played Alice, I understood Benji differently, and when I played Caden, I could see*



(a) Presence (PQ) (b) VEQ–Ownership Subscale (c) Uncanny Valley Index (UVI)

Figure 7: System experience measures across sessions. (a) Presence remained moderate and stable, (b) Embodiment was consistently moderate, and (c) Comfort increased with humanness without an uncanny valley dip.

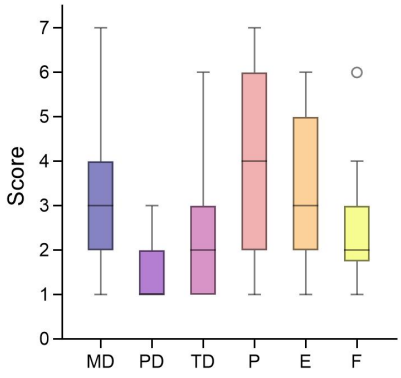


Figure 8: NASA Task Load Index (NASA-TLX) scores across the six subscales: MD, PD, TD, P, E, and F. Ratings indicate moderate mental demand and effort, low physical and temporal demand, high performance, and low frustration.

both Alice and Benji in new ways." Similarly, P07 explained that encountering avatars first as counterparts and later embodying them himself highlighted the contrast between perspectives. He reflected, *"It was interesting to see how they handled it versus how I handled it,"* noting that this shift deepened his perspective-taking across sessions. Together, these accounts suggest that role-switching fostered perspective growth within sessions and gradually consolidated empathic skills through iterative practice over time.

Avatars and Environments Supported Empathic Engagement. Avatars were frequently highlighted as important for making interactions feel natural and immersive. Participants described them as *"like real people"* (P04, P05), noting that their gestures and movements enhanced believability. P06 reflected that although the avatars resembled *"in-game characters,"* their actions nonetheless felt *"really natural."* Voices and facial animations further reinforced immersion; as P03 noted, the voices were *"pleasant,"* while expressions *"gave weight to the dialogue."* These features provided the believability cues that supported empathic engagement and sustained participants' connection to their roles throughout the training.

The surrounding virtual environment also played a critical role in fostering presence and empathy. Several participants emphasized how being *"surrounded by the environment"* (P07) made them take the task more seriously, as if they had stepped into the characters' shared world. P01 appreciated that the setting was *"very clear"* and facilitated interaction, while P04 described that seeing other roommates in VR *"made me more engaged into the conversation, like as if I was in the real world."* Such accounts suggest that the environment provided a meaningful stage for role-play, strengthening both immersion and empathic resonance.

At the same time, participants remained sensitive to the limitations of avatar nonverbal cues. P10 explained, *"in real life I can read from someone's face what they think [...] but here it was always the same face,"* and P03 noted that this limitation made him rely more on verbal cues. These reflections indicate that while avatars conveyed a range of expressions, their lack of nuance limited participants' ability to read emotions and sustain empathic engagement. This highlights the need for more refined nonverbal cues in future empathy-oriented systems to better support subtle social signaling.

4.5.3 *Qualitative Findings from Diary.*

Sixteen participants (D01–D16; $M = 23.62$, $SD = 3.03$; 11 male, 5 female) ultimately completed the entries. The thematic analysis of these diaries revealed three recurring patterns that illustrate how cognitive empathy training was reflected in daily life, through noticing differences in perspectives, adapting responses, and sustaining cognitive empathic practices beyond the VR sessions.

Recognizing Perspective Differences in Everyday Interactions. Participants reported becoming more attentive to differences in perspectives, needs, or priorities during everyday interactions. This awareness arose in social situations, ranging from casual conversations with friends to shared responsibilities or collaborative tasks, where participants reflected on how their own actions might be experienced from different standpoints. As D07 explained: *"I try to understand how others feel when they are criticized. Because*

I once criticized others for not completing their tasks in a project." Beyond immediate interactions, some participants also extended this awareness to online contexts. D09 described reflecting on polarized audience reactions to a TV series episode: *"While I personally enjoyed the latest episode, many people had strong negative reactions [. . .] I read more carefully about those negative comments that are genuine discussions with their thoughts and reasons, not just emotional outbursts. I became interested in learning others' perspectives, even if they may conflict with my own."* These accounts suggest that participants increasingly noticed and considered diverse perspectives across both everyday and mediated interactions.

Adjusting Communication and Reactions with Empathic Strategies. Beyond noticing perspective differences, participants described actively adapting their communication and reactions to foster more constructive interactions. Several noted becoming calmer and more patient in daily exchanges. As D02 reflected: *"In general I have become calmer and more understanding of other people's viewpoints. "* Similarly, D03 explained how empathic reflection helped resolve a disagreement with their partner: *"I am constantly reminded of the sessions and how important it is to check each other's perspectives before coming to a conclusion [. . .] Just today my girlfriend and I had an argument about our schedules, but we talked it through, and I think my patience has increased."* Other participants echoed this emphasis on constructive responses, highlighting greater willingness to listen and adjust how they communicated. D13 described applying these strategies in routine contexts: *"When my friends and I were discussing the dinner menu, I listened to their ideas first before questioning them. This helped me consider different perspectives and realize that my own view might be wrong."* D09 explained how they sought to balance perspectives: *"I try to get more information before making the judgement and think of whether there is room for compromising or even a win–win situation."*

These accounts illustrate how participants drew on cognitive empathy to guide their communication and responses, applying it to both personal disagreements and mundane situations, thereby fostering constructive and cooperative interactions in daily life.

Sustaining Cognitive Empathy Across Time and Contexts. Participants reported that the practices introduced in VR carried over into a wide range of daily situations, from study sessions and household negotiations to personal disagreements. For example, D04 explained: *"It helped me better to understand how different people may be, and if we need to decide something together we need to take the difference into consideration for better communication. I try to use this knowledge when I am arguing with someone."* Likewise, D06 emphasized applying perspective-taking to regulate emotional reactions: *"When I am upset about someone's behavior, I try to understand other people's perspectives [. . .] it helps in controlling impulsive reactions and showing more patience."* Beyond specific incidents, participants also highlighted the longer-term relevance of these strategies. D08 noted: *"Because they are easy to remember and they work."* D09 further linked them to relationship quality: *"Because it makes the relationship between people more harmonious."* These accounts suggest that the effects of VR training extended beyond the sessions themselves, supporting the sustained use of empathic practices across diverse contexts and over time.

At the same time, a few participants indicated that they did not perceive major changes, as they already considered themselves strong in perspective-taking. These accounts suggest that while the training fostered sustained and transferable applications of cognitive empathy for many, its impact varied depending on participants' prior dispositions.

4.6 Summary

Overall, Study 2 demonstrated that *CoEmpaTeam* effectively enhanced participants' cognitive empathy. Quantitative results showed significant early gains in both perspective-taking and fantasy. While fantasy remained consistently elevated across all training sessions, perspective-taking showed early improvement followed by session-to-session fluctuations, yet remained higher than at pre-training throughout the study. Complementary qualitative findings revealed that participants not only engaged in perspective-taking during VR interactions but also extended these practices into everyday contexts. These results suggest both immediate improvements and meaningful transfer of cognitive empathy beyond the VR setting. In the following section, we discuss the mechanisms underlying these outcomes and their implications for the design of cognitive empathy-focused training systems.

5 Discussion

In this section, we interpret our findings by discussing the mechanisms that may explain these effects, the design implications for cognitive empathy-focused training systems, and the broader relevance of our work for HCI.

5.1 How Role-Switching with LLM-Driven Avatars Cultivates Cognitive Empathy

Our findings suggest that role-switching in VR fostered cognitive empathy by engaging participants in both enactment and observation. When embodying a role such as Alice, participants not only interacted with other avatars but also observed their behaviors from a distinct standpoint. In subsequent sessions, switching into these roles required participants to reinterpret the same situation from new perspectives, for example, playing Benji after previously observing him as Alice. This process of alternating between self-as-character and observer-of-others reflects principles from social cognitive theory [3], where enactive mastery and vicarious observation jointly contribute to skill development, creating structured opportunities to practice perspective-taking.

Qualitative findings offer process-level indications of this mechanism. Participants described adopting each character's perspective, comparing these viewpoints with those they had embodied in earlier roles, and updating their understanding of the situation after switching roles. These perspective-shifting and interpretive-updating processes align with the core cognitive operations of cognitive empathy, indicating that the role-switching mechanism functioned as intended within the VR system powered by LLM-driven avatars. The stability of role identities established in Study 1 further reinforced these effects: the avatars consistently expressed coherent personalities, enabling participants to ground their reasoning in stable character traits rather than in arbitrary variation.

Further, beyond the immediate engagement, repeated switching roles across the three sessions enabled participants to revisit familiar situations from multiple angles, deepening their interpretive understanding over time. As P04 explained: *"By switching the roles, I could understand each of them more, when I played Alice, I understood Benji differently, and when I played Caden, I could see both Alice and Benji in new ways."* This cycle of enactment, observation, and reflection aligns with experiential learning theory [57], where iterative exposure transforms momentary exercises into sustained skill development.

Alongside these qualitative developments, the quantitative PT trajectory showed session-to-session fluctuations, with a temporary dip at Session 2. Research on skill acquisition indicates that rapid early gains are often followed by short-term performance variability as learners transition from initial exposure into deeper processing [62, 96]. Experiential learning theory likewise suggests that new learning situations can produce strong early engagement [57]. This aligns with the marked increase observed in Session 1, which may have been amplified by participants' limited prior VR experience. As novelty diminished and participants engaged more analytically with the task, ratings dipped temporarily rather than showing a sustained decline. Such mid-trajectory fluctuations are consistent with consolidation processes in skill development [26] and with calibration effects in which learners develop more accurate self-assessments [62]. Despite these fluctuations, PT scores remained above pre-training levels, indicating that the overall improvements in cognitive empathy were retained beyond novelty-driven effects.

In addition to cognitive gains, we observed that participants occasionally reported emotional reactions during the sessions. This is consistent with research showing that perspective-taking can indirectly evoke affective response [44], as reflecting on others' situations may elicit emotional reactions. However, the system was intentionally designed to cultivate cognitive rather than affective empathy. Following Davis's formulation [23], cognitive empathy concerns understanding another person's thoughts and motivations without requiring shared emotional states. To maintain this focus, the scenario structure, LLM-driven avatars, and role-switching in *CoEmpaTeam* were crafted to emphasize perspective-taking and reflective reasoning, processes central to cognitive empathy.

Conceptually, our findings extend prior work that emphasized narrative immersion and emotion as primary drivers of empathy [60, 92]. Participants' interactions with avatars resembled parasocial relationships [47], in which people perceive mediated characters as socially real. However, unlike traditional one-sided parasocial interaction, the LLM-driven avatars adapted their dialogue to participants' input, creating reciprocal exchanges that made the interaction feel socially responsive. This shift illustrates how VR systems can move beyond offering momentary empathic experiences to functioning as environments for cultivating cognitive empathy as a transferable skill.

Building on this distinction between passive immersion and interactive engagement, it is also important to consider how VR compares to traditional in-person role-play. While similar role-switching could, in principle, be conducted with actors or with multiple participants enacting the respective roles, our VR-based approach provides scalability, consistent scenario delivery, and adaptive interactivity. Within the same scenario, LLM-driven avatars flexibly adapt their dialogue while maintaining stable character identities, eliminating the need for multiple actors or participants. These affordances position *CoEmpaTeam* as a complement to traditional in-person training, particularly valuable when participant numbers or resources are limited. In this way, VR functions not as a replacement for human role play but as a scalable and controlled medium for structured empathy practice and transferable skill development.

5.2 Extending Cognitive Empathy Practices Beyond VR

Our findings show that cognitive empathy practices cultivated through *CoEmpaTeam* extended into participants' everyday interactions. Like prior work [92, 110], we collected immediate self-reports after each session, but we complemented this with a week-long diary that captured participants' reflections in situ. Because our evaluation was situated within a single co-living scenario, the diary served as an important complement to in-session measures by revealing how participants applied perspective-taking in daily negotiations and social encounters, and how they reflected on these experiences. This resonates with reflective learning theories [7], which highlight reflection as a bridge from situated experience to transferable practice. While such accounts cannot replace multi-scenario evaluations, they illustrate how cognitive empathy practices developed in VR may be enacted beyond the immediate training context.

Overall, these results demonstrate that empathy training outcomes were not confined to controlled sessions but also manifested in real-life contexts, thereby enhancing the ecological validity of our findings. However, it is important to distinguish short-term improvements from long-term learning effects. Prior work shows that structured perspective-taking activities can yield meaningful short-term gains in cognitive empathy [65, 91], whereas sustaining such gains typically requires continued practice, contextual reinforcement, or ongoing reflective engagement [91]. This distinction highlights a well-recognized challenge in empathy research: short-term gains often stem from situationally activated perspective-taking, whereas long-term change requires the gradual consolidation of interpretive habits across varied social contexts. Our evaluation provides preliminary evidence of short-term transfer into everyday interactions, but it does not address long-term durability or broader generalization beyond the study period. Future work should investigate how systems like *CoEmpaTeam* can be integrated into longer-term training programmes or everyday digital platforms to support sustained development over time.

5.3 Design Implications for Cognitive Empathy Training Systems

Our findings yield several design insights for the development of systems that aim to foster cognitive empathy.

5.3.1 Structured Cognitive Tasks. While immersive storytelling can make empathy training engaging [114], our results show that cognitive empathy benefits most when participants are required to reason explicitly about different viewpoints. In our case, negotiating household rules prompted them to articulate and reconcile divergent perspectives, turning immersion into active reflection. To fully

realize the potential of VR for cultivating cognitive empathy, designs should therefore move beyond immersion and narrative alone, embedding explicit cognitive interventions that make perspective-taking and critical reflection central to the experience [64].

5.3.2 Distinct and Consistent Role Design. The credibility of the training relied on avatars whose personalities and behaviors were both stable and recognizable. When participants could clearly identify a role's traits and motivations, they found it easier to distinguish between perspectives and adjust their own responses accordingly. Empathy training systems should thus prioritize character design that conveys distinct and consistent identities, expressed through personality traits, behavioral cues, and dialogue patterns, to scaffold reliable and meaningful perspective-taking.

5.3.3 Relatable Everyday Contexts. Situating the training in familiar interpersonal scenarios, such as household negotiations, made it easier for participants to apply perspective-taking strategies beyond VR. Because these scenarios mirrored the kinds of interactions they routinely encountered with roommates, friends, or classmates, participants could more readily see how empathic reflection was relevant to their own lives. This finding aligns with our initial assumption that embedding everyday conflicts into the training provides a meaningful basis for eliciting perspective-taking and, in turn, fostering the development of cognitive empathy. Empathy training systems should therefore ground practice in relatable social contexts that connect directly to users' lived experiences, increasing the likelihood that strategies rehearsed in VR transfer into everyday interactions.

Taken together, these implications suggest that VR-based cognitive empathy training benefits from structured perspective-taking tasks, persona-consistent roles, and relatable everyday contexts. In *CoEmpaTeam*, repeated role switching across three avatars supports enactment, observation, and reflection across sessions.

5.4 Scalability Considerations in Avatar Persona Design

A related systems-level consideration concerns scalability, especially in designing and validating persona-consistent avatars. In our implementation, creating coherent role profiles and ensuring behavioral consistency involved manual curation, which constrains scaling to larger deployments or rapid adaptation to new domains. For verbal behavior, recent advances in LLM-based Big Five persona shaping [69] demonstrate that linguistic style can be systematically aligned with target personality traits, offering initial steps toward reducing reliance on manually crafted personas. For nonverbal behavior, research in social signal processing has identified foundational links between expressive cues and social perceptions [108]. However, existing taxonomies are not yet sufficiently formalised or computationally actionable to support fully automated persona-consistent avatar generation. Developing such taxonomies and exploring how they might be combined with persona-shaped language models represents a promising direction for improving the scalability of multi-role systems. The modular architecture of *CoEmpaTeam*, which separates role descriptions, task structure, and dialogue generation, provides a practical foundation for incorporating these advances. As computational models of verbal and nonverbal persona expression mature, new roles or scenarios could be integrated without redesigning the overall system, reducing the dependence on human validation and enabling broader applicability.

5.5 Broader Implications for HCI

Although our study was grounded in co-living negotiations, the approach may inform a wider range of HCI contexts. At the system level, *CoEmpaTeam* suggests that LLM-driven avatars could function not only as conversational agents but also as structured partners that support perspective-taking. While we instantiated this approach in a household scenario, similar designs may be adaptable to domains such as teamwork training, intercultural communication, or healthcare education, where understanding multiple perspectives is essential. The same framework could also extend beyond cognitive empathy to other soft skills, such as conflict resolution or collaborative decision-making, by tailoring role content and task dynamics.

Our study highlights the value of pairing controlled VR training with longitudinal methods as one way to explore whether training effects extend beyond the lab. By combining in-session measures with a week-long diary, we were able to capture both immediate experiences and short-term transfer, showing how established qualitative methods can complement controlled interventions.

Finally, our study contributes to conversations about the role of empathy in HCI. Prior work has treated empathy as an experience, for example, immersing users in a designed scenario to momentarily adopt another's perspective [55]. In contrast, our findings suggest that interactive systems can serve as training environments that cultivate empathy as a transferable skill. This perspective shifts the focus from designing for one-time empathic experiences to designing for empathic capacities, equipping users with abilities that endure and can be applied across contexts and time.

5.6 Ethical Considerations

LLM-driven dialogue systems involve widely recognized ethical risks, particularly the potential to generate biased or stereotypical responses [27, 32]. In the context of cognitive empathy training, such risks are critical to address, as they can undermine the very goal of fostering perspective-taking. In *CoEmpaTeam*, avatars were intentionally crafted with distinct personalities to represent contrasting viewpoints, making behavioral differences part of the design. However, LLM-generated dialogue could still risk introducing unintended stereotypes—for instance, associating assertiveness or passivity with gender or cultural traits—that go beyond these intended characteristics. Although our study did not collect personal data and thus posed minimal privacy risks, inappropriate or exclusionary outputs could nonetheless cause discomfort. Future systems should therefore incorporate safeguards such as auditing and calibrating dialogue, as well as mechanisms for user feedback and disengagement, helping to ensure that designed role differences foster empathy without inadvertently reinforcing stereotypes.

6 Limitations and Future Work

While our findings provide evidence for the effectiveness and transfer of cognitive empathy training in VR, several limitations warrant consideration.

First, our sample was relatively homogeneous, consisting mainly of young university students (average age 24.39). This may limit the generalizability of the findings to populations, particularly other age groups or those less familiar with immersive technologies. Beyond age homogeneity, cultural differences were not examined. Although our sample included participants from multiple cultural groups, we did not analyze how cultural norms might shape role interpretation, perspective-taking strategies, or interactions with LLM-driven avatars. Given that cultural factors influence empathic reasoning and communication styles [76], future work could investigate how users from different cultural contexts engage with role-switching and whether culturally grounded adaptations of role profiles or dialogue patterns enhance the training experience. This limitation also points toward opportunities for system adaptivity. Future systems could leverage LLMs to dynamically adjust avatar characteristics to create perspective-taking challenges tailored to each user. Such adaptations could align with a user's cultural background or deliberately introduce contrasting or culturally distinct viewpoints, depending on the pedagogical goal, enabling more culturally and personally adaptive empathy-training experiences.

In addition, the study did not include a non–role-switching baseline. This aligns with our evaluation focus: *CoEmpaTeam* integrates role-switching, personality-driven LLM avatars, and embodied VR interaction into a single, cohesive training experience, and the present study examines the effectiveness of the system as a whole rather than isolating individual components. To support further research, we open-source the system to facilitate reproducibility and enable future work to develop comparative variants or baseline-controlled implementations.

Another limitation concerns technical constraints. Although advances in LLMs and speech recognition enabled fluid avatar interactions, occasional latency still disrupted conversational flow. Prior work suggests that delays longer than about four seconds can negatively impact user experience [78]. Future systems could adopt established strategies to mitigate this issue, such as incorporating hesitation gestures (e.g., touching the chin) or verbal fillers (e.g., short reflective phrases) to simulate human-like pauses and preserve conversational continuity.

Our evaluation approach also presents limitations, as it relied primarily on self-reports, which capture subjective experiences but may not fully reflect the complexity of empathic processes. Complementary biosignals, such as heart rate variability, respiration, or galvanic skin response, could provide insights into how cognitive empathy unfolds during and after training [68, 88]. Beyond serving as measurement tools, biosignals also hold potential as expressive media in social VR. Because avatars' nonverbal cues, such as facial expressions, are often less nuanced than in real life, integrating biosignal-based feedback (e.g., visualizing heart rate or breathing rhythm through avatars) could offer an additional channel for conveying affective and cognitive states, supporting more natural and empathic communication [68]. Building on these possibilities, future extensions of *CoEmpaTeam* could explore affective aspects of empathy more directly, for example by incorporating avatars capable of expressing affective cues and by leveraging biosignal-based visualization to enrich emotional expressivity in VR.

Finally, our role design was limited to a co-living scenario with three predefined avatars. While this served as a proof of concept, the framework can be extended to roles representing diverse demographics, cultural backgrounds, or even physical conditions, allowing users to embody perspectives further removed from their own [117]. In future work, supporting customizable role configurations and scenario design could enhance engagement and adaptability, enabling the system to extend beyond empathy training toward the cultivation of broader soft skills, such as conflict resolution, collaborative decision-making, and other transferable competencies. Overall, these directions highlight opportunities for *CoEmpaTeam* to evolve from the current implementation into a flexible platform for fostering transferable skills across diverse domains.

7 Conclusion

We introduce *CoEmpaTeam*, a novel VR-based system that leverages LLM-driven avatars with distinct personalities and a structured role-switching mechanism to facilitate cognitive empathy training. In Study 1, we validated through avatar self-assessment and human evaluation that the three designed avatars reliably expressed their intended personalities, providing a credible foundation for subsequent cognitive empathy training. In Study 2, we showed that repeated role-switching improved participants' perspective-taking and fantasy scores, with the acquired skills transferring into real-world contexts as captured in a diary study. Collectively, these findings highlight the potential of structured role-switching in VR to cultivate cognitive empathy, and we hope they can inform future HCI research on designing interactive systems that foster cognitive empathy skills.

References

[1] Setareh Aghel Manesh, Tianyi Zhang, Yuki Onishi, Kotaro Hara, Scott Bateman, Jiannan Li, and Anthony Tang. 2024. How people prompt generative ai to create interactive vr scenes. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*. 2319–2340.

[2] Ruth S Aylett, Sandy Louchart, Joao Dias, Ana Paiva, and Marco Vala. 2005. FearNot!–an experiment in emergent narrative. In *International workshop on intelligent virtual agents*. Springer, 305–316.

[3] Albert Bandura et al. 1986. Social foundations of thought and action. *Englewood Cliffs, NJ* 1986, 23-28 (1986), 2.

[4] Philippe Bertrand, Jérôme Guegan, Léonore Robieux, Cade Andrew McCall, and Franck Zenasni. 2018. Learning empathy through virtual reality: multiple strategies for training empathy-related abilities using body ownership illusions in embodied virtual reality. *Frontiers in Robotics and AI* 5 (2018), 326671.

[5] Sarah Lynne Bowman. 2014. Educational live action role-playing games: A secondary literature review. *The Wyrd Con Companion Book* 3 (2014), 112–131.

[6] Jeremy Boy, Anshul Vikram Pandey, John Emerson, Margaret Satterthwaite, Oded Nov, and Enrico Bertini. 2017. Showing people behind data: Does anthropomorphizing visualizations elicit more empathy for human rights data?. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 5462–5474.

[7] Evelyn M Boyd and Ann W Fales. 1983. Reflective learning: Key to learning from experience. *Journal of humanistic psychology* 23, 2 (1983), 99–117.

[8] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[9] Chantal MLR Brazeau, Robin Schroeder, Sue Rovi, and Linda Boyd. 2010. Relationships between medical student burnout, empathy, and professionalism climate. *Academic medicine* 85, 10 (2010), S33–S36.

[10] Iago Alves Brito, Julia Soares Dollis, Fernanda Bufon Färber, Pedro Schindler Freire Brasil Ribeiro, Rafael Teixeira Sousa, et al. 2025. Integrating Personality into Digital Humans: A Review of LLM-Driven Approaches for Virtual Reality. *arXiv preprint arXiv:2503.16457* (2025).

[11] Lauren Buck, Gareth W Young, and Rachel McDonnell. 2023. Avatar customization, personality, and the perception of work group inclusion in immersive virtual reality. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*. 27–32.

[12] Linnda R Caporael. 1997. The evolution of truly social cognition: The core configurations model. *Personality and Social Psychology Review* 1, 4 (1997), 276–298.

[13] Justine Cassell. 2001. Embodied conversational agents: representation and intelligence in user interfaces. *AI magazine* 22, 4 (2001), 67–67.

[14] Antonella Cavallaro, Marco Romano, and Rossana Laccone. 2024. Examining User Perceptions to Vocal Interaction with AI Bots in Virtual Reality and Mobile Environments: A Focus on Foreign Language Learning and Communication Dynamics. In *International Conference on Human-Computer Interaction*. Springer, 20–30.

[15] Doungmani Chongruksa, Penprapa Prinyapol, Yuhamasaulaet Wadeng, and Chaiwat Padungpong. 2010. Storytelling: program for multicultural understanding and respect among Thai-Buddhist and Thai-Muslim students. *Procedia-Social and Behavioral Sciences* 5 (2010), 282–288.

[16] Frederik Roland Christiansen, Linus Nørgaard Hollensberg, Niko Bach Jensen, Kristian Julsgaard, Kristian Nyborg Jespersen, and Ivan Nikolov. 2024. Exploring presence in interactions with llm-driven npcs: A comparative study of speech recognition and dialogue options. In *Proceedings of the 30th ACM Symposium on Virtual Reality Software and Technology*. 1–11.

[17] Federico Cioffi, Massimiliano Masullo, Aniello Pascale, and Luigi Maffei. 2025. Speech Intelligibility in Virtual Avatars: Comparison Between Audio and Audio–Visual-Driven Facial Animation. In *Acoustics*, Vol. 7. MDPI, 30.

[18] Vicky Clark, Keith Tuffin, and Natilene Bowker. 2020. Managing conflict in shared housing for young adults. *New Zealand Journal of Psychology* 49, 1 (2020), 4–13.

[19] Alison N Cooke, Doris G Bazzini, Lisa A Curtin, and Lisa J Emery. 2018. Empathic understanding: Benefits of perspective-taking and facial mimicry instructions are mediated by self-other overlap. *Motivation and emotion* 42, 3 (2018), 446–457.

[20] Wenzhe Cui, Rui Liu, Zhi Li, Yifan Wang, Andrew Wang, Xia Zhao, Sina Rashidian, Furqan Baig, IV Ramakrishnan, Fusheng Wang, et al. 2023. Glancewriter: Writing text by glancing over letters with gaze. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–13.

[21] Max T Curran, Jeremy Raboff Gordon, Lily Lin, Priyashri Kamlesh Sridhar, and John Chuang. 2019. Understanding digitally-mediated empathy: An exploration of visual, narrative, and biosensory informational cues. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[22] Mary Czerwinski, Eric Horvitz, and Susan Wilhite. 2004. A diary study of task switching and interruptions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 175–182.

[23] Mark H Davis. 1983. Measuring individual differences in empathy: evidence for a multidimensional approach. *Journal of personality and social psychology* 44, 1 (1983), 113.

[24] Mark H Davis et al. 1980. A multidimensional approach to individual differences in empathy. (1980).

[25] María del Pilar Sánchez-López and Virginia Dresch. 2008. The 12-Item General Health Questionnaire (GHQ-12): reliability, external validity and factor structure in the Spanish population. *Psicothema* 20, 4 (2008), 839–843.

[26] K Anders Ericsson and Andreas C Lehmann. 1996. Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual review of psychology* 47, 1 (1996), 273–305.

[27] Anna Fang, Hriday Chhabria, Alekhya Maram, and Haiyi Zhu. 2025. Social Simulation for Everyday Self-Care: Design Insights from Leveraging VR, AR, and LLMs for Practicing Stress Relief. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–23.

[28] Mariana Fernandez-Espinosa, Kara Clouse, Dylan Sellars, Danny Tong, Michael Bsales, Sophonie Alcindor, Timothy D Hubbard, Michael Villano, and Diego Gómez-Zará. 2025. Breaking the Familiarity Bias: Employing Virtual Reality Environments to Enhance Team Formation and Inclusion. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–16.

[29] Lucy Foulkes, Ariyana Reddy, Juliette Westbrook, Elizabeth Newbronner, and Dean McMillan. 2019. The impact of housemate relationships on undergraduate student wellbeing. (2019).

[30] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. 2015. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 1631–1640.

[31] Adam D Galinsky, Gillian Ku, and Cynthia S Wang. 2005. Perspective-taking and self-other overlap: Fostering social bonds and facilitating social coordination. *Group processes & intergroup relations* 8, 2 (2005), 109–124.

[32] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics* 50, 3 (2024), 1097–1179.

[33] Uğur Genç and Himanshu Verma. 2024. Situating Empathy in HCI/CSCW: A Scoping Review. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–37.

[34] Thorsten Gieser. 2008. Embodiment, emotion and empathy: A phenomenological approach to apprenticeship learning. *Anthropological theory* 8, 3 (2008), 299–318.

[35] Alex Gillespie and Beth Richardson. 2011. Exchanging social positions: Enhancing perspective taking within a cooperative problem solving task. *European journal of social psychology* 41, 5 (2011), 608–616.

[36] Lewis R Goldberg. 2013. An alternative "description of personality": The Big-Five factor structure. In *Personality and personality disorders*. Routledge, 34–47.

[37] Mar Gonzalez-Franco and Tabitha C Peck. 2018. Avatar embodiment. towards a standardized questionnaire. *Frontiers in Robotics and AI* 5 (2018), 74.

[38] Xiang Gu, Sheng Li, Kangrui Yi, Xiaojuan Yang, Huiling Liu, and Guoping Wang. 2022. Role-exchange playing: An exploration of role-playing effects for anti-bullying in immersive virtual environments. *IEEE transactions on visualization and computer graphics* 29, 10 (2022), 4215–4228.

[39] Manuel Guimarães, Rui Prada, Pedro A Santos, João Dias, Arnav Jhala, and Samuel Mascarenhas. 2020. The impact of virtual reality in the social presence of a virtual agent. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–8.

[40] Insook Han, Hyoung Seok Shin, Yujung Ko, and Won Sug Shin. 2022. Immersive virtual reality for increasing presence and empathy. *Journal of Computer Assisted Learning* 38, 4 (2022), 1115–1126.

[41] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[42] Robert Hassan. 2020. Digitality, virtual reality and the 'empathy machine'. *Digital journalism* 8, 2 (2020), 195–212.

[43] Simone Hausknecht, Michelle Vanchu-Orosco, and David Kaufman. 2019. Digitising the wisdom of our elders: Connectedness through digital storytelling. *Ageing & Society* 39, 12 (2019), 2714–2734.

[44] Meghan L Healey and Murray Grossman. 2018. Cognitive and affective perspective-taking: evidence for shared and dissociable anatomical substrates. *Frontiers in neurology* 9 (2018), 491.

[45] Fernanda Herrera, Jeremy Bailenson, Erika Weisz, Elise Ogle, and Jamil Zaki. 2018. Building long-term empathy: A large-scale comparison of traditional and virtual reality perspective-taking. *PloS one* 13, 10 (2018), e0204494.

[46] Chin-Chang Ho and Karl F MacDorman. 2010. Revisiting the uncanny valley theory: Developing and validating an alternative to the Godspeed indices. *Computers in Human Behavior* 26, 6 (2010), 1508–1518.

[47] Donald Horton and R Richard Wohl. 1956. Mass communication and para-social interaction: Observations on intimacy at a distance. *psychiatry* 19, 3 (1956), 215–229.

[48] Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2023. PersonaLLM: Investigating the ability of large language models to express personality traits. *arXiv preprint arXiv:2305.02547* (2023).

[49] Karin Johansson, Raquel Robinson, Jon Back, Sarah Lynne Bowman, James Fey, Elena Márquez Segura, Annika Waern, and Katherine Isbister. 2024. Why Larp? A Synthesis Article on Live Action Roleplay in Relation to HCI Research and Practice. *ACM Transactions on Computer-Human Interaction* 31, 5 (2024), 1–35.

[50] Hyojin Ju, Jungeun Lee, Seungwon Yang, Jungseul Ok, and Inseok Hwang. 2025. Toward Affective Empathy via Personalized Analogy Generation: A Case Study on Microaggression. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–31.

[51] Gary Kacmarcik. 2006. Using natural language to manage NPC dialog. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 2. 115–117.

[52] Molly Kelly Grealy, Emmet Godfrey, Finn Brady, Erin Whyte O'Sullivan, Grace A Carroll, and Tom Burke. 2022. Borderline personality disorder traits and mentalising ability: The self-other social cognition paradox. *Frontiers in psychiatry* 13 (2022), 1023348.

[53] Robert S Kennedy, Norman E Lane, Kevin S Berbaum, and Michael G Lilienthal. 1993. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology* 3, 3 (1993), 203–220.

[54] Nari Kim, Sangsu Jang, Hansol Kim, Jaeyeon Lee, and Young-Woo Park. 2023. Design and Field Trial of Tunee in Shared Houses: Exploring Experiences of Sharing Individuals' Current Noise-level Preferences with Housemates. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[55] Sameer Kishore, Bernhard Spanlang, Guillermo Iruretagoyena, Shivashankar Halan, Dalila Szostak, and Mel Slater. 2019. A virtual reality embodiment technique to enhance helping behavior of police toward a victim of police racial aggression. *PRESENCE: Virtual and Augmented Reality* 28 (2019), 5–27.

[56] Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 453–456.

[57] David A Kolb. 2014. *Experiential learning: Experience as the source of learning and development*. FT press.

[58] Sara H Konrath, William J Chopik, Courtney K Hsing, and Ed O'Brien. 2014. Changes in adult attachment styles in American college students over time: A meta-analysis. *Personality and Social Psychology Review* 18, 4 (2014), 326–348.

[59] Annett Körner, Michael Geyer, Marcus Roth, Martin Drapeau, Gabriele Schmutzer, Cornelia Albani, Siegfried Schumann, and Elmar Brähler. 2008. Personality assessment with the NEO-five-factor inventory: the 30-item-short-version (NEO-FFI-30). *Psychotherapie, Psychosomatik, Medizinische Psychologie* 58, 6 (2008), 238–245.

[60] Martijn JL Kors, Erik D Van Der Spek, Julia A Bopp, Karel Millenaar, Rutger L Van Teutem, Gabriele Ferri, and Ben AM Schouten. 2020. The curious case of the transdiegetic cow, or a mission to foster other-oriented empathy through virtual reality. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

[61] Leon OH Kroczek, Alexander May, Selina Hettenkofer, Andreas Ruider, Bernd Ludwig, and Andreas Mühlberger. 2025. The influence of persona and conversational task on social interactions with a LLM-controlled embodied conversational agent. *Computers in Human Behavior* (2025), 108759.

[62] Justin Kruger and David Dunning. 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology* 77, 6 (1999), 1121.

[63] Anita K Kuhnley, Tram H Nguyen, Alexandra C Gantt, and Patricia Hinkley. 2023. Creatively increasing empathy: The impacts of an online empathy workshop. *Journal of Creativity in Mental Health* 18, 1 (2023), 60–72.

[64] Eugene Kukshinov, Federica Gini, Anchit Mishra, Nicholas Bowman, Brendan Rooney, and Lennart E Nacke. 2025. Seeing is not thinking: Testing capabilities of VR to promote perspective-taking. *IEEE Transactions on Visualization and Computer Graphics* (2025).

[65] Tony Chiu Ming Lam, Klodiana Kolomitro, and Flanny C Alamparambil. 2011. Empathy training: Methods, evaluation practices, and validity. *Journal of Multidisciplinary Evaluation* 7, 16 (2011), 162–200.

[66] Ray Land. 2004. Issues of embodiment and risk in online learning. In *Beyond the comfort zone: Proceedings of the 21st ASCILITE Conference*. 530–538.

[67] Jaewook Lee, Jun Wang, Elizabeth Brown, Liam Chu, Sebastian S Rodriguez, and Jon E Froehlich. 2024. GazePointAR: A context-aware multimodal voice assistant for pronoun disambiguation in wearable augmented reality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–20.

[68] Sueyoon Lee, Abdallah El Ali, Maarten Wijntjes, and Pablo Cesar. 2022. Understanding and designing avatar biosignal visualizations for social virtual reality entertainment. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.

[69] Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona Diab, and Maarten Sap. 2025. Big5-chat: Shaping llm personalities through training on human-grounded data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 20434–20471.

[70] Ziming Li, Pinaki Prasanna Babar, Mike Barry, and Roshan L Peiris. 2024. Exploring the use of large language model-driven chatbots in virtual reality to train autistic individuals in job communication skills. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.

[71] Ziming Li, Pinaki Prasanna Babar, and Roshan L Peiris. 2025. Generative role-play communication training in virtual reality for autistic individuals: A study on job coach experiences in vocational training programs. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–22.

[72] Ziyi Liu, Zhengzhe Zhu, Lijun Zhu, Enze Jiang, Xiyun Hu, Kylie A Peppler, and Karthik Ramani. 2024. Classmeta: Designing interactive virtual classmate to promote vr classroom participation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–17.

[73] Jose Llanes-Jurado, Lucía Gómez-Zaragozá, Maria Eleonora Minissi, Mariano Alcañiz, and Javier Marín-Morales. 2024. Developing conversational virtual humans for social emotion elicitation based on large language models. *Expert Systems with Applications* 246 (2024), 123261.

[74] Tracey Loughran, Kate Mahoney, and Daisy Payling. 2022. Women's voices, emotion and empathy: engaging different publics with 'everyday'health histories. *Medical Humanities* 48, 4 (2022), 394–403.

[75] Zexin Ma. 2020. Effects of immersive stories on prosocial attitudes and willingness to help: testing psychological mechanisms. *Media Psychology* 23, 6 (2020), 865–890.

[76] Hazel Rose Markus and Shinobu Kitayama. 2014. Culture and the self: Implications for cognition, emotion, and motivation. In *College student development and academic life*. Routledge, 264–293.

[77] Daniel Martin, Sandra Malpica, Diego Gutierrez, Belen Masia, and Ana Serrano. 2022. Multimodality in VR: A survey. *ACM Computing Surveys (CSUR)* 54, 10s (2022), 1–36.

[78] Mykola Maslych, Mohammadreza Katebi, Christopher Lee, Yahya Hmaiti, Amirpouya Ghasemaghaei, Christian Pumarada, Janneese Palmer, Esteban Segarra Martinez, Marco Emporio, Warren Snipes, et al. 2025. Mitigating Response Delays in Free-Form Conversations with LLM-powered Intelligent Virtual Agents. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*. 1–15.

[79] Dan P McAdams. 2001. The psychology of life stories. *Review of general psychology* 5, 2 (2001), 100–122.

[80] Robert R McCrae and Oliver P John. 1992. An introduction to the five-factor model and its applications. *Journal of personality* 60, 2 (1992), 175–215.

[81] Mikko Meriläinen. 2012. The self-perceived effects of the role-playing hobby on personal development: A survey report. *International Journal of Role-Playing* 3 (2012), 49–68.

[82] Daphne A Muller, Caro R Van Kessel, and Sam Janssen. 2017. Through Pink and Blue glasses: Designing a dispositional empathy game using gender stereotypes and Virtual Reality. In *Extended abstracts publication of the annual symposium on computer-human interaction in play*. 599–605.

[83] Aline Normoyle, João Sedoc, and Funda Durupinar. 2024. Using llms to animate interactive story characters with emotions and personality. In *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 632–635.

[84] Süeda Özkaya, Santiago Berrezueta-Guzman, and Stefan Wagner. 2025. How LLMs are Shaping the Future of Virtual Reality. *arXiv preprint arXiv:2508.00737* (2025).

[85] Mengxu Pan, Alexandra Kitson, Hongyu Wan, and Mirjana Prpa. 2025. Ellma-t: an embodied llm-agent for supporting english language learning in social vr. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*. 576–594.

[86] Xueni Pan and Antonia F de C Hamilton. 2018. Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape. *British Journal of Psychology* 109, 3 (2018), 395–417.

[87] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*. 1–22.

[88] Elena Parra Vargas, Aitana García Delgado, Sergio C Torres, Lucía A Carrasco-Ribelles, Javier Marín-Morales, and Mariano Alcañiz Raya. 2022. Virtual reality stimulation and organizational neuroscience for the assessment of empathy. *Frontiers in Psychology* 13 (2022), 993162.

[89] Frederic W Platt and Vaughn F Keller. 1994. Empathic communication: a teachable and learnable skill. *Journal of General Internal Medicine* 9, 4 (1994), 222–226.

[90] Zhongfei Qing, Zhongang Cai, Zhitao Yang, and Lei Yang. 2023. Story-to-motion: Synthesizing infinite and controllable character animation from long text. In *SIGGRAPH Asia 2023 technical communications*. 1–4.

[91] Helen Riess, John M Kelley, Robert W Bailey, Emily J Dunn, and Margot Phillips. 2012. Empathy training for resident physicians: a randomized controlled trial of a neuroscience-informed curriculum. *Journal of general internal medicine* 27, 10 (2012), 1280–1286.

[92] Mohammad Rashidujjaman Rifat, Reem Ayad, Ashratuz Zavin Asha, Bingjian Huang, Selin Okman, Dina Sabie, Hasan Shahid Ferdous, Robert Soden, and Syed Ishtiaque Ahmed. 2024. Cohabitant: The design, implementation, and evaluation of a virtual reality application for interfaith learning and empathy building. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–19.

[93] Aitor Rovira, David Swapp, Bernhard Spanlang, and Mel Slater. 2009. The use of virtual reality in the study of people's responses to violent incidents. *Frontiers in behavioral neuroscience* 3 (2009), 917.

[94] Gery W Ryan and H Russell Bernard. 2003. Techniques to identify themes. *Field methods* 15, 1 (2003), 85–109.

[95] Thomas Schäfer and Marcus A Schwarz. 2019. The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in psychology* 10 (2019), 813.

[96] Richard A Schmidt and Robert A Bjork. 1992. New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological science* 3, 4 (1992), 207–218.

[97] Neal Schmitt. 1996. Uses and abuses of coefficient alpha. *Psychological assessment* 8, 4 (1996), 350.

[98] Eva Sentas, John M Malouff, Bernadette Harris, and Caitlin E Johnson. 2018. Effects of teaching empathy online: A randomized controlled trial. *Scholarship of Teaching and Learning in Psychology* 4, 4 (2018), 199.

[99] Chenxinran Shen, Joanna Mcgrenere, and Dongwook Yoon. 2024. LegacySphere: Facilitating intergenerational communication through perspective-taking and storytelling in embodied VR. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–16.

[100] Donghee Shin. 2018. Empathy and embodied experience in virtual environment: To what extent can virtual reality stimulate empathy and embodied experience? *Computers in human behavior* 78 (2018), 64–73.

[101] Alon Shoa, Ramon Oliva, Mel Slater, and Doron Friedman. 2023. Sushi with einstein: Enhancing hybrid live events with llm-based virtual humans. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*. 1–6.

[102] Adam Smith. 2010. *The theory of moral sentiments*. Penguin.

[103] John Sweller. 2011. Cognitive load theory. In *Psychology of learning and motivation*. Vol. 55. Elsevier, 37–76.

[104] Jason Tham, Ann Hill Duin, Laura Gee, Nathan Ernst, Bilal Abdelqader, and Megan McGrath. 2018. Understanding virtual reality: Presence, embodiment, and professional practice. *IEEE Transactions on Professional Communication* 61, 2 (2018), 178–195.

[105] Endel Tulving et al. 1972. Episodic and semantic memory. *Organization of memory* 1, 381-403 (1972), 1.

[106] Austin Van Loon, Jeremy Bailenson, Jamil Zaki, Joshua Bostick, and Robb Willer. 2018. Virtual reality perspective-taking increases cognitive empathy for specific others. *PloS one* 13, 8 (2018), e0202442.

[107] Susanne Van Mulken, Elisabeth André, and Jochen Müller. 1998. The persona effect: How substantial is it?. In *People and computers XIII: Proceedings of HCI'98*. Springer, 53–66.

[108] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and Vision Computing* 27, 12 (2009), 1743–1759. doi:10.1016/j.imavis.2008.11.007 Visual and multimodal analysis of human spontaneous behaviour:.

[109] Hongyu Wan, Jinda Zhang, Abdulaziz Arif Suria, Bingsheng Yao, Dakuo Wang, Yvonne Coady, and Mirjana Prpa. 2024. Building llm-based ai agents in social virtual reality. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.

[110] R Michael Winters, Bruce N Walker, and Grace Leslie. 2021. Can you hear my heartbeat?: hearing an expressive biosignal elicits empathy. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–11.

[111] Bob G Witmer and Michael J Singer. 1998. Measuring presence in virtual environments: A presence questionnaire. *Presence* 7, 3 (1998), 225–240.

[112] Shu-I Wu, Shen-Ing Liu, Yih-Jer Wu, Ling-Lang Huang, Thih-ju Liu, Kai-Liang Kao, and Yu-Hsia Lee. 2023. The efficacy of applying the Interpersonal Effectiveness skills of dialectical behavior therapy into communication skills workshop for clinical nurses. *Heliyon* 9, 3 (2023).

[113] Yuanjie Wu, Yu Wang, Sungchul Jung, Simon Hoermann, and Robert W Lindeman. 2021. Using a fully expressive avatar to collaborate in virtual reality: Evaluation of task performance, presence, and attraction. *Frontiers in Virtual Reality* 2 (2021), 641296.

[114] Yao Xu, Ding Ding, Yongxin Chen, Zhuying Li, and Xiangyu Xu. 2024. iStray-Paws: Immersing in a Stray Animal's World through First-Person VR to Bridge Human-Animal Empathy. In *Proceedings of the 30th ACM Symposium on Virtual Reality Software and Technology*. 1–11.

[115] Nick Yee and Jeremy Bailenson. 2007. The Proteus effect: The effect of transformed self-representation on behavior. *Human communication research* 33, 3 (2007), 271–290.

[116] Ruoxin You, Yihao Zhou, Weicheng Zheng, Yiran Zuo, Mayra Donaji Barrera Machuca, and Xin Tong. 2023. BlueVR: Design and evaluation of a virtual reality serious game for promoting understanding towards people with color vision deficiency. *Proceedings of the ACM on Human-Computer Interaction* 7, CHI PLAY (2023), 289–318.

[117] Kexin Zhang, Edward Glenn Scott Spencer Jr, Abijith Manikandan, Andric Li, Ang Li, Yaxing Yao, and Yuhang Zhao. 2025. Inclusive avatar guidelines for people with disabilities: Supporting disability representation in social virtual reality. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–26.

[118] Ke Zhao, Ruiqi Chen, Xiaziyu Zhang, Chenxi Wang, Siling Chen, Xiaoguang Wang, Yujue Wang, and Xin Tong. 2025. Immersive Biography: Supporting Intercultural Empathy and Understanding for Displaced Cultural Objects in Virtual Reality. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.

[119] Xiuqi Tommy Zhu, Heidi Cheerman, Minxin Cheng, Sheri R Kiami, Leanne Chukoskie, and Eileen McGivney. 2025. Designing VR Simulation System for Clinical Communication Training with LLMs-Based Embodied Conversational Agents. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–9.