VITA: A Multi-Modal LLM-Based System for Longitudinal, Autonomous and Adaptive Robotic Mental Well-Being Coaching

MICOL SPITALE, Politecnico di Milano, Milan, Italy and University of Cambridge, Cambridge, United Kingdom of Great Britain and Northern Ireland
MINJA AXELSSON, Department of Computer Science and Technology, University of Cambridge, Cambridge, United Kingdom of Great Britain and Northern Ireland
HATICE GUNES, Department of Computer Science and Technology, University of Cambridge, Cambridge, United Kingdom of Great Britain and Northern Ireland

Recently, several works have explored if and how robotic coaches can promote and maintain mental well-being in different settings. However, findings from these studies revealed that these robotic coaches are not ready to be used and deployed in real-world settings due to several limitations that span from technological challenges to coaching success. To overcome these challenges, this article presents VITA, a novel multi-modal LLM-based system that allows robotic coaches to autonomously adapt to the coachee's multi-modal behaviours (facial valence and speech duration) and deliver coaching exercises in order to promote mental well-being in adults. We identified five objectives that correspond to the challenges in the recent literature, and we show how the VITA system addresses these via experimental validations that include one in-lab pilot study ($N = 4$) that enabled us to test different robotic coach configurations (pre-scripted, generic and adaptive models) and inform its design for using it in the real world, and one real-world study ($N = 17$) conducted in a workplace over 4 weeks. Our results show that: (i) coachees perceived the VITA adaptive and generic configurations more positively than the pre-scripted one, and they felt understood and heard by the adaptive robotic coach, (ii) the VITA adaptive robotic coach kept learning successfully by personalising to each coachee over time and did not detect any interaction ruptures during the coaching and (iii) coachees had significant mental well-being improvements via the VITA-based robotic coach practice. The code for the VITA system is openly available via https://github.com/Cambridge-AFAR/VITA-system.

CCS Concepts: • **Human-centered computing** → *Empirical studies in HCI*; **HCI theory, concepts and models**;

Additional Key Words and Phrases: mental well-being, robotic coach, open-source system, reinforcement learning, adaptation, autonomous, human-robot interaction

This work was undertaken and finalised while M. Spitale was a postdoctoral researcher at the University of Cambridge. M. Spitale and H. Gunes have been supported by the EPSRC/UKRI under grant EP/R030782/1 (ARoEQ). M. Axelsson is funded by the Osk. Huttunen Foundation and the EPSRC under grant EP/T517847/1.
Authors' Contact Information: Micol Spitale (corresponding author), Politecnico di Milano, Milan, Italy and University of Cambridge, Cambridge, United Kingdom of Great Britain and Northern Ireland; e-mail: micol.spitale@polimi.it; Minja Axelsson, Department of Computer Science and Technology, University of Cambridge, Cambridge, United Kingdom of Great Britain and Northern Ireland; e-mail: minjaaxelsson@gmail.com; Hatice Gunes, Department of Computer Science and Technology, University of Cambridge, Cambridge, United Kingdom of Great Britain and Northern Ireland; e-mail: Hatice.Gunes@cl.cam.ac.uk.

ACM 2573-9522/2025/3-ART38
https://doi.org/10.1145/3712265

ACM Reference format:
Micol Spitale, Minja Axelsson, and Hatice Gunes. 2025. VITA: A Multi-Modal LLM-Based System for Longitudinal, Autonomous and Adaptive Robotic Mental Well-Being Coaching. *ACM Trans. Hum.-Robot Interact.* 14, 2, Article 38 (March 2025), 28 pages.
https://doi.org/10.1145/3712265

1 Introduction

Mental health and well-being promotion and maintenance have been paramount goals for the World Health Organization[1] in the last decade, given the increase of mental health conditions worldwide. Despite the rise of mental healthcare needs, the treatment coverage remains extremely low leading to a call for innovative mental healthcare solutions. To address the limited access to healthcare systems, several works [32, 38, 61] have recently investigated if and how robotic coaches can be used to promote mental well-being in different contexts, e.g., workplace [57], student accommodations [30] and so on. These studies show that the robotic technologies have the potential of providing a dedicated space and time for a coaching session that smartphone-based apps cannot [6]. Findings from these studies revealed that robotic coaches are still very far from being used and deployed in the real world as fully autonomous coaching agents for long-term use due to several limitations that span from the technological challenges to the success of the practice.

Overall, findings in [30, 57] showed that coachees expect from the robotic coaches more than they can currently deliver. The coachees complained about the lack of interactivity and responsiveness of the robotic coach [6, 57] or the lags in robot response and behaviour and various glitches [32] (*Limitation 1*, *L1*). Findings from past works also showed that coachees expected the robotic coach to 'adapt and personalise more' [57]. However, none of the past works have explored the use of personalised and adaptive robotic coaching deploying it in the long term despite the recommendations distilled by [7] and [32] (*L2*). Additionally, coachees experienced several **Interaction Ruptures** (**IRs**; i.e., coachees felt awkward and/or the robot was making a mistake, e.g., interrupting the coachees [59]) that may have also jeopardized the coach–coachee alliance, another important factor for the success of the coaching practice [15] (*L3*). Coachees also highlighted the importance of the conversational capabilities of the robotic coaches. They reported that the dialogue with the robot was very limited [32] and that they did not feel understood or heard [57] (*L4*). Past works have demonstrated the efficacy of robotic coaching (i.e., leading to an improvement of coachees' mental well-being) when delivering coaching exercises in short-term interactions (e.g., after 7-days interaction [30]). The very few papers [9, 57] that explored their efficacy in the long term did not find any significant improvements in the coachees' mental well-being (*L5*). The factors that could have negatively influenced the success of the practice may be multiple, including the technical limitations of the robots highlighted above.

In this article, we present *VITA* ('life' in Latin), a novel multi-modal **Large Language Model (LLM)**-based system for adaptive mental well-being coaching that allows a robotic coach to autonomously adapt to each coachee and deliver coaching exercises to promote mental well-being in adults by addressing the abovementioned limitations as reported in Table 1. The main objectives of this work are the following. First, we design and develop the robotic coach to be interactive (i.e., multi-modal two-way interaction between the coach and the coachee) and responsive (i.e., the robotic coach can respond to the coachee by analysing both the coachees' verbal and non-verbal behaviours), and we evaluate the coachees' perceptions towards the robotic coach (*Objective 1*, *O1* addressing L1). Second, we develop and implement an adaptive and personalised model to manage

[1] https://www.who.int/health-topics/mental-health

Table 1. Limitations of Past Works (L), Objectives to Address Those Limitations (O) and Experimental Validation (EV) of the VITA System Designed to Achieve Those Objectives (EV)

Limitations of past works (L)	**Objectives of this work (O)**	**EV**
(L1) Coachees reported negative perception and feedback about the robotic coach because they found it *not* interactive nor responsive [59] or 'glitchy' [32]	*(O1)* Making the robotic coach more interactive and responsive to improve coachees' perception towards the robotic coach	Negative Attitude Towards Robots Scale (NARS) questionnaire and Game Experience Questionnaire (GEQ) results of the pilot study (see Section 4, *EV1.1*) and the real-world study (see Section 5.5.2, *EV1.2*), and coachees's interview comments on the perception of the robotic coach (see Section 5.5.3, *EV1.3*)
(L2) The robotic coach practice was *not personalised* longitudinally, in contrast to recommendations of [7] and [32] and coachees expected it to 'adapt and personalise more' [57]	*(O2)* Implementing and evaluating an adaptive robotic coach to better match with the coachee's behaviour over time	Reward increasing over time in the pilot study (see Figure 6, *EV2.1*) and in the real-world study (see Figure 7 and Section 5.5.1, *EV2.2*), coachees' perception of adaptation (see Section 5.5), their overall impression (see Table 4, *EV2.3*) and coachees's interview comments on the perception of the robotic coach (see Section 5.5.3, *EV2.4*)
(L3) Coachees experienced *several interaction ruptures* during robotic coaching [59] that can negatively affect the coach–coachee alliance	*(O3)* Recognising the occurrences of interaction ruptures via automatic detection during robotic coaching to guarantee the success of the practice	Interaction ruptures detected in the real-world study (see Section 5.5.2, *EV3.1*), and coachees's interview comments on the flow of the coaching practice (see Section 5.5.3, *EV3.2*)
(L4) Coachees found the conversational capabilities of the robotic coach were limited, as in [30, 57], and they did not feel understood or listened to	*(O4)* Embedding the robot with the advance large language model (LLM) module to imitate human coach behaviours	Coachees's interview comments on conversational capabilities of the robotic coach (see Section 5.5.3, *EV4.1*)
(L5) Coachees' mental well-being *did not significantly* improve via robotic coaching in the long term [57]	*(O5)* *Significant* improvement in the coachees' mental well-being to demonstrate the success of the robotic coaching	*Significant* Ryff's Psychological Well-Being Scale (RPWS) scale results in the real-world study (see Figure 7 and Section 5.5.2, *EV5.1*), and coachees's interview comments on the benefits of the coaching practice (see Section 5.5.3, *EV5.2*)

the dialogue flow of the coaching practice and evaluate its advantages by comparing it with a non-adaptive model (*O2* addressing L2).

Then, we implement an automatic model based on machine learning to recognise IRs during the interaction with the robotic coach, and we embed it into the robotic coaching system to evaluate whether ruptures occur during the coaching practice (*O3* addressing L3). We chose to analyse *IRs* specifically, as we examined these in our previous work addressing barriers to robotic well-being coaching [59]. Within robotic well-being coaching and in consultation with a human coach/therapist, IRs were defined as instances of a user expressing awkwardness and/or the robot making a mistake in coaching [59]. The relationship between the coach and coachee is important to coaching success [15], and difficulties in this relationship can pose barriers to successful coaching [11]. Such difficulties include the coachee not feeling supported by the coach, or the coach not being involved, sensitive or flexible during coaching [11]. Such awkward coaching interactions may make a user feel embarrassed, nervous or self-conscious [34], posing a barrier to coaching. As such, both robot mistakes and user awkwardness are important barriers to well-being coaching [59], contrary to research of robots in other contexts, which focuses on only the mistakes or communication failures of the robot (e.g., [26]).

We then leverage the advancements in LLMs and embed a ChatGPT-based conversational module using prompt engineering into the robotic coach system to deliver positive psychology exercises (*O4* addressing L4). Finally, we evaluate if coachees who interacted in the long term with a fully autonomous and adaptive robotic coach reported significant improvements in their mental well-being (*Objective 5*, *O5* addressing L5).

To simultaneously address all these objectives in a single work, we conducted various evaluations by undertaking a pilot in-lab study—that involved four participants—to evaluate the differences between *pre-scripted* (i.e., the robotic coach follows a pre-defined sequence of utterances to speak aloud to the coachee no matter what the coachee says), *generic* (i.e., the robotic coach understands via natural language processing what the coachee says and responds accordingly without personalising the selection of the next line of the dialogue to coachee's behaviour) and *adaptive* (i.e., the robotic coach understands via natural language processing what the coachee says and personalises to the coachee's behaviour by selecting the next line of the dialogue flow accordingly) models. This initial study informed us that the adaptive model led to a more personalised interaction than the pre-scripted and generic models. Based on this finding, we embedded the robotic coach system with the adaptive model to manage the dialogue flow during the coaching practice delivered by the robot. We deployed our implementation in a long-term real-world study—that involved 17 participants at a tech company—to evaluate the capabilities of the VITA-based robotic well-being coaching.

The main contributions of this article are two-fold. First, we provide the community with a novel and open source system named VITA to design and develop an autonomous and adaptive robotic coach that can deliver mental well-being practices by adapting to each coachee. Our second contribution is the evaluation of the VITA-based robotic coach via pilot and long-term real-world studies to successfully deliver mental well-being coaching practices. Specifically, we found that (i) coachees perceived the adaptive and generic configurations more positively than the pre-scripted one, and they felt understood and heard by the robotic coach, (ii) the adaptive robotic coach kept learning successfully by personalising to each coachee over time and did not detect any IRs during the coaching practice and (iii) coachees reported significant mental well-being improvements via the robotic coach practice.

2 Related Work

2.1 Robotic Coaches for Mental Well-Being

Coaching is a practice focused on promoting the well-being of mentally healthy people, for instance to aid them in reaching new levels of achievement and success in their professional or personal life. Coaching does not aim to address or treat mental illness, contrary to psychological therapy [24]. The goals of coaching might include improving the coachee's ambition, morale, or overall well-being [21]. Techniques differ across types of coaching. For instance, Positive Psychology aims to aid the coachee to focus on the positive experiences and sensations in their life [52], while Cognitive Behavioural Coaching focuses on the relationship between a coachee's thoughts, emotions and behaviour [21]. The success of coaching depends on the relationship between the coach and the coachee, relying on factors such as mutual understanding and trust [15] as well as transparency in the coaching relationship [23]. Previously, the quality of therapeutic relationships has been examined through the alliance between the therapist and client (here, analogous to coach and coachee). Alliance has been investigated through the bond between the therapist and client, as well as how the goals of the therapeutic process are met [41].

Robots are promising tools for well-being coaching due to their advantages over other technological tools, such as mobile applications. Additionally, past research exploring robots for mindfulness meditation found that robots provide the benefit of creating a *dedicated time and space* for well-being practice [6], which using a mobile application for the same purpose would not. In a past study exploring robotic coaches for positive psychology exercises at a workplace, participants remarked that the robot acted as a *visual reminder* of the well-being practice they had practiced with it earlier [57]. In another study, participants were found to self-disclose more and attain more emotional benefits from a journal writing exercise with a robot, in comparison to a voice agent [51]. In comparison to human coaching, while robots cannot perform at a human level, robots can be more accessible and available [5]. In the case of our study, having a shared robot at a workplace could provide low-barrier access to well-being coaching for employees.

Just a handful of papers have researched the use of robotic coaches to foster mental well-being, e.g., [1, 6, 9, 32, 38, 53, 57, 61]. Jeong et al. [30] conducted a longitudinal study, which utilized Jibo robots to provide positive psychology interventions to students in home settings over a period of seven days and reported that participants experienced a better sense of well-being, improved mood and a readiness to change, while gradually developing an affinity for the robot. Shi et al. [53] investigated the effect of physical embodiment and personalisation on the user-perceived quality of **Text-to-Speech (TTS)** voices for mindfulness. Their results showed that the user-personalized TTS voices were able to perform nearly as well as human voices, indicating that user personalisation could be a powerful approach to raise user perception of TTS voice quality. Spitale et al. [57, 59] conducted a study involving employees of a tech company to interact with two different forms of robotic coaches that delivered positive psychology exercises over 4 weeks. Their results showed that the robot form may impact the perception of the coachees towards the robotic coach. Jeong et al. [31] explored the companion-like robotic coach behaviour during well-being therapy by comparing the assistant, coach and companion roles of the robot. Their results showed that the companion robot was the most effective in building a positive therapeutic alliance with its users.

2.2 Human-Robot Interaction (HRI) Tools and Systems

There are notable efforts within the HRI community to create tools for composing social human–robot interactions [20, 47, 55], including robot-specific tools, (e.g., Choreograph [46]), content authoring tools (e.g., [14]) and dialogue management tools (e.g., [28]). These tools are not multi-modal, open source, robot-agnostic, composable and modular—crucial factors for adoption and wide usage by the HRI community [60]. One representative example that addresses the aforementioned open issues is HARMONI [60]—an open source, robotic-agnostic, modular and multi-modal tool to help researchers compose social human–robot interactions. Such tools are extremely beneficial when maintained and supported by the HRI community as they enable researchers to quickly implement a new interaction or deploy a new robot for a research study without wasting time to reinvent the wheel.

Such tools are usually implemented for generic use. On the contrary, robotic *systems* (e.g., [44]) refer to the development and, if applied, the use of the abovementioned tools, to specific application scenarios, such that they ‘synthesise underlying techniques to achieve system-level HRI behaviour’[2]. To this aim, Taylor and Riek [64] presented the Robot-Centric Group Detection and Tracking System as a new system that enables robots to detect and track groups of using a crowd-aware, tracking by-detection approach that has been made publicly available. Analogously, Nanavati et al. [42] introduced a novel system that relies on informed direction selection to avoid

[2] https://humanrobotinteraction.org/2024/fullpaper/

obstacles and traverse a hallway, and periodic human help to charge, evaluating it in the wild and shared their code open source. However, none of the existing robotic systems have focussed on robotic well-being coaching. This article presents a novel adaptive robotic system named VITA that leverages the HARMONI tool to promote code reuse and advances the current HRI state of the art by enabling the autonomous delivery of well-being practices.

2.3 Adaptation and Personalisation in HRI

Within the HRI literature, very recently, a number of works, e.g., [4, 13, 19, 40], have started to investigate and evaluate the use of adaptive and personalised robots by showing that adaptive configurations are very promising for robotic applications [49]. For example, Axelsson and Skantze [4] presented a fully automated system for building adaptive presentations for embodied agents. They evaluated the system involving 43 participants who interacted with the adaptive system, and their results demonstrated that the user preferred the adaptive system. Analogously, Gillet et al. [19] explored how learning robot gaze behaviours (via **Reinforcement Learning (RL)**) can balance human participation in conversational interactions. Their results showed that the proposed reward for the RL approach enabled the robot to encourage participants to take more turns. Only a few works have explored the applications of adaptation and personalisation in robots in the context of mental well-being. Churamani et al. [13] proposed a novel system for well-being coaching by utilising continual learning to personalise the robotic coaching to each user and compared static, adaptive and personalised versions of the robotic coach. Their results showed that overall the users tend to prefer the robotic coach with continual personalisation.

However, none of the previous works have created *long-term autonomous adaptation* and/or *personalisation* strategies for *robotic mental well-being coaching*.

3 VITA System

This section introduces the VITA system providing a general overview and a detailed description of the main components embedded in the robotic coach that have been used and tested during the pilot and real-world studies to address the objectives listed in Table 1.

3.1 Overview

VITA is based on a *behaviour tree* structure, which runs at a frequency of 10 Hz and is managing all the VITA components from a high level. The behaviour tree, inspired from Spitale et al. [60]—in which the authors presented a very simple pre-scripted dialogue-based flow—coordinates reading the data from the sensors (microphone and camera of the robot), running the detectors continuously (e.g., facial expression, IRs), calling external services (e.g., OpenAI API for natural language processing) and activating the actuators (e.g., speaker and motors of the robot) when a decision about the dialogue flow is made.

A walk-through of the interaction is depicted in Figure 1 and described as follows. To begin, the sensors (microphone and camera) and the detectors (openSMILE, OpenFace, facial expression, voice activity and *IRs*) are activated. The *dialogue flow manager* initiates the interaction, and the robot introduces the positive psychology exercise and asks the first question. Then, the coachee is required to answer the questions posed by the robotic coach and the detectors analyse the raw audio and video data published by the microphone and the camera of the robot. After that, the detectors send their outputs to the dialogue flow manager which decides, on the basis of the multi-modal behavioural data collected from the detectors, which dialogue action to take next in the coaching practice and sends the corresponding prompt to the *ChatGPT engine*. This produces its textual response and sends it to the actuators that synthesise the text and generate the robotic coach behaviour (i.e., sound, facial lip-sync and gesture).

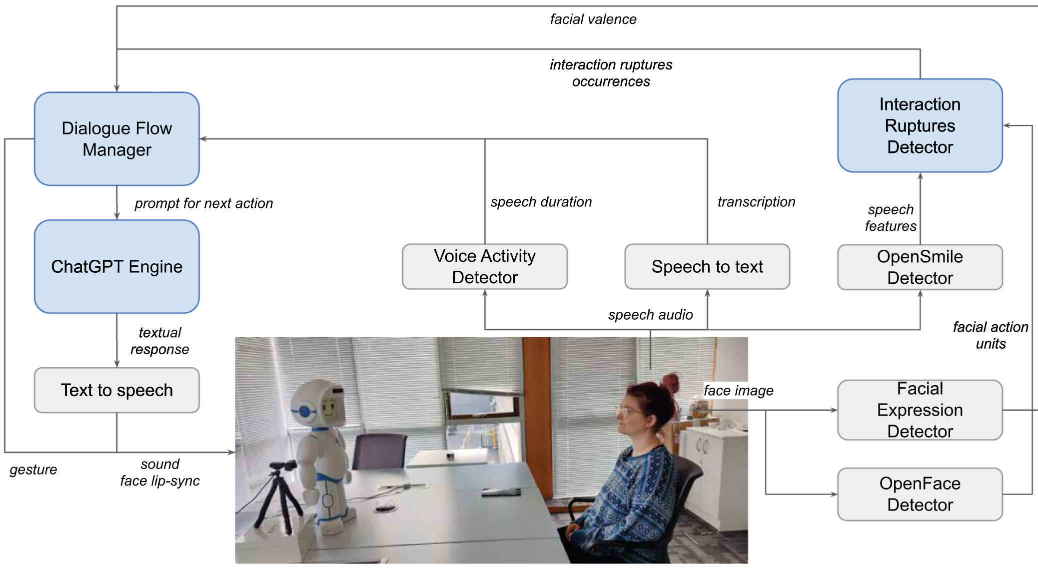


Fig. 1. The components of the VITA system for adaptive robotic mental well-being coaching.

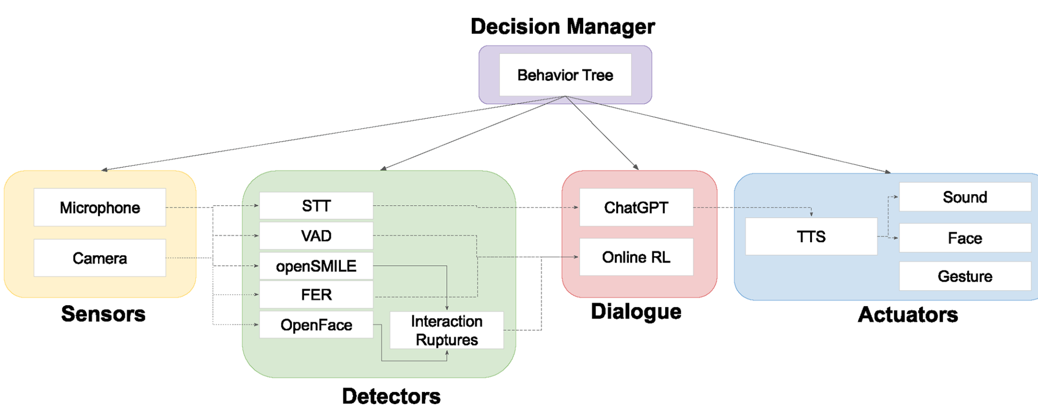


Fig. 2. Architecture of the HARMONI modules integrated in the robotic coach system.

3.2 VITA Modules

We designed the robotic coach system architecture by including sensor, detector, actuator, dialogue and decision modules using the open source framework HARMONI [60]. Figure 2 represents the modules integrated in the robotic coach via the HARMONI framework. We describe each of these modules in the following sections.

3.2.1 Detector Modules. Besides relying on the robot's sensors (microphone and camera), we embedded in the robot the following detector modules to perceive the environment and the behavioural cues of the coachee.

— Voice activity detector: We used the VAD library[3] to detect when the coachee is speaking using the microphone input.

[3] https://github.com/marsbroshok/VAD-python

— Facial **Action Unit (AU)** detector: We integrated the OpenFace[4] library for real-time detection of facial action unit features.
— Speech feature detector: We integrated openSMILE[5] library for real-time extraction of speech features, e.g., pitch.
— Facial expression recognition detector: We integrated the FaceChannel library[6] to extract the valence and arousal from coachees' facial gestures.
— IR detector: We embedded the IR detector (described in Section 3.3).
— Speech to text module: We used the Google speech-to-text API[7] to transcribe the speech of the coachees.

3.2.2 Actuator Modules. Besides the wrapper for the robot motor, we embedded the robot with actuators to make the interaction more natural.

— Facial expression actuator: We used the CordialFace [54] to control the facial expressions of the robot and display the lip-sync on the face screen.
— TTS actuator: We used Amazon Polly TTS API[8] to synthesize the speech of the robot. We picked Amy voice as in [57].

3.2.3 Dialogue Modules. We implemented the following dialogue modules.

— Natural language processing module: We used chatGPT to analyse the coachees' responses by integrating the OpenAI API[9] in the robotic system.
— ADAPT-RL: We integrated the trained Adapt-RL model reported in Section 3.2 to make sequential decisions on which action to take next for delivering the positive psychology exercises.

3.2.4 Decision Manager Module. We implemented a high-level decision module as a behaviour tree by using the py-tree library[10] to manage the whole interaction.

3.3 Dialogue Flow of Adaptive Coaching

In previous studies [6, 57], participants noted that the robotic coach was not interactive and responsive enough due to the pre-scripted nature of the interaction (*L1*). Also, coachees revealed that they would expect the robotic coach to adapt and personalise more (*L2*). VITA attempts to overcome these limitations by providing a system that enables fully autonomous and adaptive robotic coaching to improve the perceptions of the coachees towards the robotic coach (*O1*) and emulate the behaviour of a human coach (*O2*). To this end, a robotic coach needs to be trained with human coach data. Hence, we collected a dataset of human–human dyadic interactions during positive psychology practices as the existing literature lacks such a dataset.

3.3.1 ***Human–Human Dyadic Interactions For Positive Psychology (HHI4PP)*** *Dataset.* We collected a dataset of human–human dyadic interactions between a human well-being coach and five participants (coachees), named HHI4PP. The coachees were three females and two males 29–33 years old ($M = 30.6$ years old, $SD = 2.19$ years old) who work as researchers at the Affective Intelligence and Robotics Lab, University of Cambridge. The human well-being coach delivered

[4] https://github.com/TadasBaltrusaitis/OpenFace
[5] https://www.audeering.com/research/opensmile/
[6] https://github.com/pablovin/FaceChannel
[7] https://cloud.google.com/speech-to-text
[8] https://aws.amazon.com/polly/
[9] https://platform.openai.com/
[10] https://py-trees.readthedocs.io/en/devel/

four different positive psychology exercises (about 15 minutes each). The exercises were the same as in our previous works [57] and [59]. The exercises were: (1) *savouring*, where a participant is asked to recall and reflect on a positive memory in the recent past [56]; (2) *gratitude*, where a participant is asked to recall instances where they felt grateful [22]; and (3) *accomplishments*, where a participant is asked to think of recent accomplishments and the positive qualities in themselves that helped them accomplish them [22]. The fourth exercise used in previous work (i.e., optimism about the future) was substituted with the *one door closes, one door opens* exercise. In this exercise, a participant is asked to think about instances where they missed out on an opportunity (i.e., a door closed) and what opportunities arose as a result (i.e., a door opened) [33]. This exercise focuses on cultivating optimism about the future. We substituted the exercise used in past studies because participants in our previous work [57] found the previous exercise particularly difficult (as reported in [59]).

The audio–visual recordings of the sessions were done via MS Teams, and we transcribed the speech using the automatic MS Teams transcriber. Our dataset includes a total of 5.5 hours of video recordings with the corresponding transcriptions of coach–coachee dyadic interactions. The ethics committee of the University of Cambridge approved the collection of this dataset, and the participants signed an informed consent before taking part in this study.

3.3.2 RL Problem Formulation. From the HHI4PP dataset, we observed (by sorting through the videos and labelling the actions) that the human well-being coach adopted a specific conversational flow to deliver the positive psychology practice, namely after first asking about an experience or a memory recall (e.g., a savouring memory) they decided whether to ask a follow-up question, summarise what the coachee just shared, or ask for a new episode (i.e., a new experience or memory). One of the main limitations of previous works [30, 57] was the pre-scripted nature of the coaching dialogue flow that made the interaction very repetitive and not responsive. Hence, we formulated the problem of conducting the well-being practice in HRI as a sequential decision-making problem for selecting the next action in the *coaching dialogue flow*. At every turn t (i.e., the robotic coach asks questions and the participant answers) [19], the robotic coach environment is captured as an observation state variable s_t. The robotic coach, similar to the human well-being coach, may choose an action a_t at every turn t that allows it to move forward with the conversation for delivering coaching practice by asking for a follow-up question, summarising what the coachee just said, or asking for a new episode. The main goal is to learn a conversational policy π: $s_t \rightarrow a_t$ that enables the successful delivery of the well-being practice (i.e., the robotic coach can deliver successfully the positive psychology exercise to promote coachee mental well-being). We formulated this as a batch (or *generic*) RL problem—i.e., the RL model was pre-trained by using the dataset described in Section 3.3.1—and *adaptive* RL—where the generic pre-trained RL model kept learning during the interaction with the coachees. We used the HHI4PP dataset because it includes audio–visual recordings and transcriptions of a human coach delivering a positive psychology practice, from which our model could learn how to make a decision on the coaching dialogue flow.

We created an 11-element vector, called s_t, to represent the *observation state* for the coaching dialogue flow, which consists of: prediction of IR (present or absent), current well-being exercise (savouring, gratitude, accomplishment, one door closes one door opens), speech features (duration of speech and silence) and previous actions (summarisation, follow-up question and new episode). All of these features were collected at the end of each turn t to keep track of the dialogue flow and the conversational interchange between the human coachee and the robotic coach, as in [66]. The actions a_t were three discrete dialogue *actions* of the robotic coach that can decide the coaching dialogue flow of the well-being practice, namely (1) summarise what the coachee said, (2) ask for a follow-up question (e.g., 'How does this event make you feel?') and (3) ask to share another episode (e.g., 'Can you share with me another thing you were grateful for during the last week?'). We defined the *reward* as the behavioural cues of the coachees, specifically as the combination of speech duration and facial valence. We decided to use the facial valence as a component of the reward since it carries information about the emotional state of the coachee [2] as also was done in a previous work [40], while we used the speech duration because we observed from the HHI4PP dataset that the human coach decided the coaching flow according to the amount of information shared by the coachee. We computed the facial valence similarly to McQuillin et al. [40], using deviations in valence values determined from coachees' facial expressions compared to their baseline facial expression (we refer to this as FV_t).

We followed the same procedure for the speech duration by normalizing the values of speech duration (SD_t). The final reward was the combination of these two values as follows:

$$R[s_t] = FV_t + SD_t \tag{1}$$

The distribution of the reward of the HHI4PP is characterised by a mean value equal to -2.72, an SD equal to *5.59* and a median value equal to -3.09. The total number of samples is *748*.

3.3.3 RL Models. We trained our batch (*generic*) RL model using three different approaches: **Deep Q-Learning (DQN)**, double DQN and neural fitted Q iteration algorithm to learn the conversational flow of a positive psychology exercise, using the d3rlpy library.[11] We opted for DQN models due to their effectiveness in handling increased complexity in the number of states and actions. These models efficiently adapt to the randomness of the environment (as in-the-wild human–robot interaction), outperforming simpler Q-learning models in this regard [18]. The results showed that there were not significant differences among the three models. We then chose the DQN as it is the most commonly used one in discrete problems, e.g., [66].

Our *adaptive* RL model uses the pre-trained *generic* RL model and is designed to be fine-tuned with real-time data of each coachee interacting with the robotic coach as depicted in Figure 3. We created personalised models (i.e., one for each coachee) that could further learn and adapt to the specific coachee behaviour over time (i.e., across multiple interactions). We picked the DQN model to pre-train our *adaptive* RL model and to learn in real time using the same d3rlpy library. This adaptive approach differs from the generic one because it keeps learning online from the data during the interaction with each coachee as in Figure 3.

3.4 Automatic Detection of IRs

Findings from a previous work [59] showed that several IRs may occur during a coachee–coach interaction (*L3*). Motivated by these findings, we decided to embed the robotic coach with the capability to detect *IRs* (i.e., the user expression of awkwardness and/or reactions to the robot mistakes [59]) (*O3*). As discussed in Section 1, the term *IRs* is related to the 'deterioration in the quality of the relationship between patient and coach' [50] which can be a barrier to coaching. Such ruptures can occur when, for example, the robot makes mistakes during the coaching session which might negatively impact the alliance and trust perceived by the user towards the robot, which may lead to an unsuccessful well-being practice [59].

3.4.1 Dataset and Labels. In [57], a robotic positive psychology coach has been deployed at a workplace for over 4 weeks. The robotic coach followed a pre-scripted interaction and it conducted four positive psychology exercises, described in detail in Spitale et al. [57]. We requested the collected dataset from [59], which contains the extracted audio–visual features of 100 videos of about 10 minutes each that were annotated in terms of occurrences of IRs with binary labels

[11] https://d3rlpy.readthedocs.io/

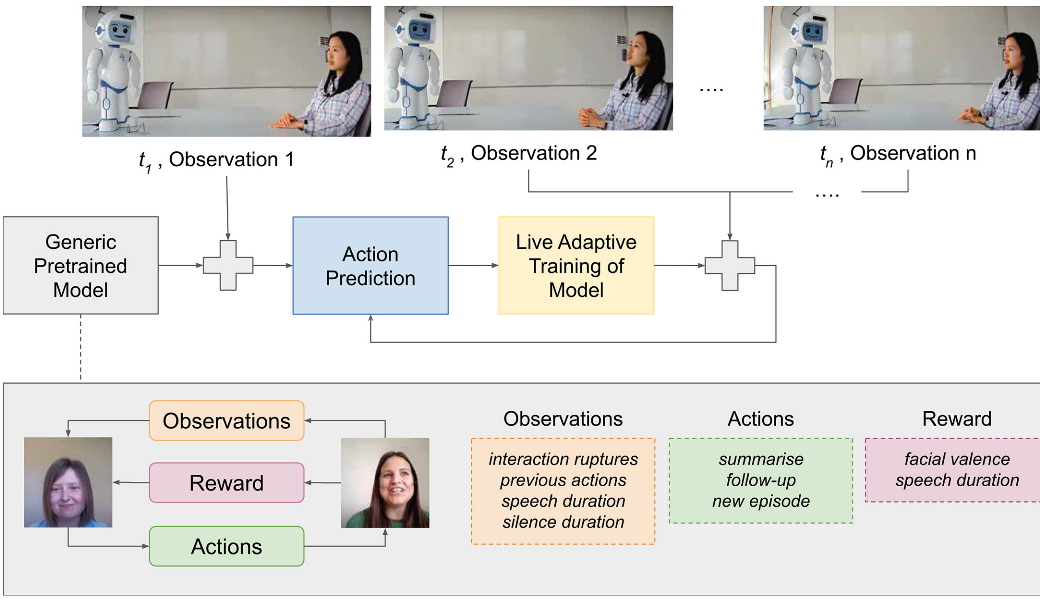


Fig. 3. RL models: generic pre-trained model (HHI4PP dataset) and the adaptive RL model fine-tuned with the real-time coachee data.

(i.e., 1: present, or 0: absent), as described in [59]. The dataset [59] includes for each video clip: (i) 35 temporal facial features extracted using the OpenFace 2.2.0 toolkit [8], such as AU1 (inner brow raiser), AU2 (outer brow raiser), AU4 (brow lowerer), AU5 (upper lid raiser) and so on; and (ii) 25 temporal interpretable speech features using the openSMILE toolbox [17], such as loudness, pitch, length of the coachees' silence and speech and so on. The dataset also includes body features, but we discarded them since they have been found to not contribute to the detection of IRs [59].

3.4.2 Feature Pre-Processing. We used the sequences of both audio and visual features with the purpose of learning temporal patterns in audio and facial cues to predict *IRs*. To this end, we created the input samples for our models by resampling all feature sequences at intervals of 1 second to capture the temporal variability of the data. We then applied to both feature sets a sliding window approach with a window size of 10*s* (chosen after running a set of experiments with 5-, 10-, 15- and 20-seconds time window) and an overlap of 3*s* between sequential windows. Given the unbalanced nature of the dataset [59], we decided to under-sample our feature sets using the NearMiss method that selected examples based on the distance of majority class examples to minority class examples. Our final tensors have the following dimensions (number of samples, window size and features): audio feature set (3602, 10, 25) and facial feature set (3602, 10, 35).

3.4.3 Automatic Detection. We formulated the IR detection as a sliding-window sequence-based binary classification problem, to detect whether an IR was present or not (No-IR) in a 10-seconds time window of an audio–visual segment (refer to Figure 4 for the model pipeline). We experimented with three temporal deep learning models: **Long Short-Term Memory (LSTM)** networks [25], **Gate Recurrent Unit (GRU)** networks [16] and **Bi-Directional LSTM (Bi-LSTM)** networks. We conducted a 5-fold stratified cross-validation, repeated 10 times (50 folds in total) with a subject-independent approach (i.e., the samples of the same subject were included either in the training or test set but not both). We chose stratified cross-validation to maintain the same proportion of 'IR' and 'No-IR' participants in each fold. We also tuned the hyperparameters on the

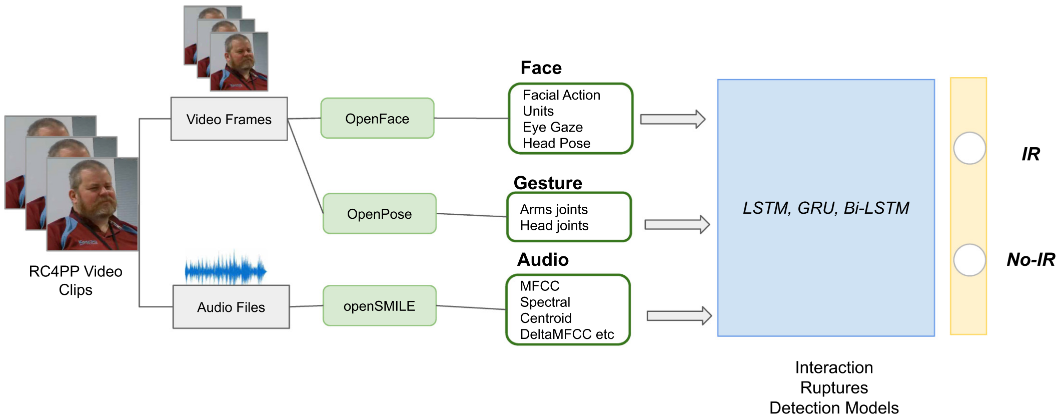


Fig. 4. The model pipeline for detecting interaction ruptures using the dataset collected in [59].

training sets with the Optuna framework [3]. Within each cross-validation experiment, all features in the training and testing set of each fold were standardized using the feature distributions of the training set via z-normalization, as in [39].

We conducted multiple experiments to compare the uni-modal versus multi-modal approaches as in [12]. We first trained our models using either facial or audio features, and then, we adopted two fusion strategies (either early or late fusion) to experiment with the multi-modal approaches, i.e., including audio and facial features as in [12]. The early fusion strategy consisted of concatenating features from the audio and facial modalities that resulted in a single vector of features, while the late fusion strategy consisted of majority voting, namely the final decision is made according to the classifier whose predicted class probability is the highest across the different uni-modal models [12].

3.4.4 Evaluation Metrics. The evaluation metrics computed for each of the 50 cross-validation folds included: (1) accuracy (ACC), the number of correct predictions with respect to the total number of predictions in the test set; (2) precision, the percentage of 'IR' samples among all samples that were classified as ruptures; (3) recall, the percentage of samples accurately identified as 'IRs' among all 'IR' samples in the dataset and (4) F1, the combination of the precision and recall of the classifier (by taking their harmonic mean), as in most classification problems [27].

These metrics were inspected across cross-validation folds and used to examine the performances of various modelling strategies. Our problem is an event detection task similar to [35], where the model should detect IRs using a sliding-window approach in an HRI context. In this scenario, as in [39], the precision of the IR detection is crucial (e.g., a robot relying on precise IR detection to appropriately repair those ruptures) to minimise false positives and maintain coaching effectiveness. For this reason, our primary metric for identifying the highest performing model was *precision*. We prioritised precision rather than the precision-recall tradeoff. Discussions with a human coach led to our decision that predicting false IRs and making repairs based on this can be very costly in a coaching practice.

3.4.5 Results. Table 2 reports the results obtained from the experiments for uni-modal and multi-modal modelling of LSTM, GRU and Bi-LSTM models. Our results show that the Bi-LSTM late fusion approach achieved an average precision of 81%, substantially higher than the average precision of other models (around 66%). Given that precision is our primary metric, the late fusion multi-modal Bi-LSTM approach has been chosen as the model to embed in our robotic coaching system.

Table 2. Experimental Results for Uni-Modal and Multi-Modal Approaches Using Facial and Audio Features across Three Deep Learning Models (LSTM, Bi-LSTM and GRU) Averaged across the 50 Folds

Model	ACC	Precision	Recall	F1
Facial				
LSTM	0.58 ± 0.10	0.62 ± 0.08	0.58 ± 0.10	0.56 ± 0.11
Bi-LSTM	0.59 ± 0.10	0.63 ± 0.10	0.59 ± 0.10	0.57 ± 0.10
GRU	0.58 ± 0.10	0.63 ± 0.09	0.58 ± 0.10	0.57 ± 0.11
Audio				
LSTM	**0.73 ± 0.09**	0.77 ± 0.07	**0.73 ± 0.09**	**0.73 ± 0.09**
Bi-LSTM	**0.73 ± 0.10**	0.78 ± 0.08	**0.73 ± 0.10**	**0.73 ± 0.10**
GRU	0.70 ± 0.10	0.77 ± 0.08	0.72 ± 0.10	0.72 ± 0.09
Facial + Audio (Early Fusion)				
LSTM	0.54 ± 0.09	0.52 ± 0.10	0.52 ± 0.09	0.51 ± 0.11
Bi-LSTM	0.55 ± 0.09	0.58 ± 0.10	0.55 ± 0.09	0.52 ± 0.11
GRU	0.55 ± 0.09	0.52 ± 0.10	0.53 ± 0.09	0.51 ± 0.11
Facial + Audio (Late Fusion)				
LSTM	0.67 ± 0.10	0.80 ± 0.09	0.53 ± 0.18	0.62 ± 0.16
Bi-LSTM	0.68 ± 0.12	**0.81 ± 0.16**	0.53 ± 0.19	0.63 ± 0.17
GRU	0.68 ± 0.11	0.78 ± 0.15	0.52 ± 0.10	0.63 ± 0.18

Bold values refer to the best results for each evaluation metric.

3.5 Prompt Engineering of LLM

To achieve a more natural and multi-modal interaction and embed the robotic coach with more sophisticated conversational capabilities (*L4*), the robotic coach has been designed to analyse the coachees' speech content and use an LLM (i.e., ChatGPT) to understand what the coachee says and respond appropriately to imitate a human coaching practice (*O4*). We have used the LLM with all the precautions needed by implementing several safety measures to avoid any unexpected or inappropriate responses as recommended in [7] and [10]. First, we integrated a natural language processing layer (i.e., moderation APIs from OpenAI) to identify whether the coachee's request content is inappropriate (e.g., self-harm, sexual) and whether the LLM's answers are inappropriate. If the layer detected an inappropriate request, the robotic coach would answer 'I found your answer very inappropriate. I would stop here the coaching practice and call the researcher' and stop the interaction. Second, we conducted several adversarial testings to better understand how the LLM replies to controversial requests (e.g., the coachee shares that they have punched someone at work), see Section B of the Appendix for more examples. Finally, we have implemented 'Prompt engineering' to constrain the topic and tone of the output text. This aims to reduce the possibility of producing undesired content, even if a coachee would try to produce it. Table 3[12] summarises the prompts for each action of the robotic coach. Note that the inclusion of the moderation API layer caused a lag in the response of the robot, but we thought it was essential for a safe and ethical conversation with the robotic coach.

We have used *gpt-3.5-turbo* model to request to 'complete' the conversation between the coachee and the robotic coach. We fed the model with the system context (i.e., the description of the exercise that the robotic coach has to deliver), and then, we incorporated the coachee's transcriptions. We

[12]Note that gpt-3.5-turbo takes as input 'human' and 'AI' conversational utterances.

Table 3. Prompt Engineering for Making a Completion Request to ChatGPT in a Coaching Practice

Action	**Prompt Engineering**
(1) Summarising	Can you please summarise what the Human has just shared?
(2) Asking for a follow-up question	Can you please ask me a follow-up question about the exercise episode I have just shared?
(3) New episode	Can you please ask me about a new episode to share?

repeated the same prompting (appending one of the prompts reported in Table 3 according to the decision made by the robotic coach as described in Section 3.3) for the whole coaching practice. We kept also the history of the dialogue in the memory of the LLM in order to provide the LLM model with the context of the conversational coaching practice. An example of the conversation is reported in Appendix A.

The ultimate goal of VITA is to understand whether it can enable a significant improvement of mental well-being not only limited to short-term interactions (*L5*) but also in the long-term and in real-world settings (*O5*).

3.5.1 LLM-Based Ethical Lessons Learned. Using LLMs to generate content for a robotic coach requires several important considerations prior to and during an empirical human–robot interaction study. We share below our experiences and recommendations in order to guide the researchers that intend to use the VITA system in their own research works.

Firstly, prior to the study, a thorough ethics review of the empirical study was undertaken by our departmental ethics review committee at the Department of Computer Science and Technology, University of Cambridge. They reviewed and approved the study design, the experiment protocol and the consent forms. We recommend that researchers considering using the VITA system should engage with the appropriate ethics committee at their institution and discuss with them whether its use is appropriate and what safety measures should be taken.

Secondly, we implemented several safety measures withing the system while using ChatGPT (as detailed in Section 3.5): adding an NLP layer to detect inappropriate user generated language, conducting adversarial testing for controversial requests (see Appendix B) and 'prompt engineering' to constrain the topic and tone of the output. These measures were taken to keep the well-being interaction appropriate and on topic. We recommend that researchers using the VITA system also undertake such careful measures and particularly consider what kinds of adversarial tests are relevant for their intended use and context and which prompts appropriately guide and constrain the intended interaction.

Thirdly, to further ensure that the interaction was kept within the boundaries of an appropriate well-being interaction session, we scaffolded the aforementioned engineered prompts within a structured well-being interaction with pre-scripted robot utterances (see Appendix A for details). This decision aligns with our previous work on design and ethical recommendations for robotic well-being coaches [7], where professional human practitioners recommended that 'verbal adaptation should be limited to preserve well-being practice efficacy'. By scaffolding LLM-generated (i.e., verbally adapted) sentences within a structured interaction, we aimed to apply the appropriate level of adaptation for this well-being interaction, while also reducing the possibility of the interaction extending outside the context of a well-being interaction, and minimizing the possibility of misunderstandings via inappropriately generated utterances. We recommend that researchers using the VITA system consider what level of verbal adaptation is appropriate and safe for the intended interaction and apply structured scaffolding as appropriate.

Finally, we recruited a study population that we pre-screened for anxiety and depression (as discussed in Section 5.1). We chose this approach in order to safeguard our participants prior to the interaction (as also recommended in our prior work [7]). We consider it ethical to thoroughly examine and investigate well-being technologies with populations without mental health difficulties, in order to mitigate potential risks that may arise with vulnerable populations. We recommend that researchers evaluate how appropriate using the VITA system is for their intended application context and whether their population includes users who may be vulnerable to potential discomfort arising in the interaction or who might need additional safety measures.

We do urge researchers using the VITA system in the future to carefully consider all of these recommendations. This is not to say that merely following or considering these ethical recommendations is sufficient for an approach, interaction, or application area to be 'ethical'. There is limited prior work on the ethics of implementing LLM models in embodied systems, particularly for well-being. Future research is needed to thoroughly investigate these issues—especially prior to applying such technology outside of a controlled and safely monitored research environment.

4 Pilot Study

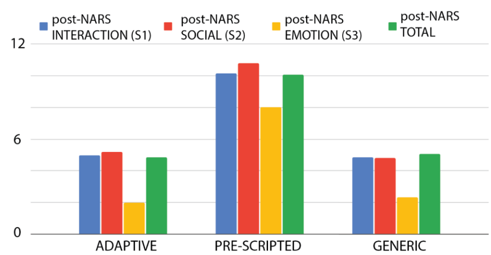
This section describes the in-the-lab pilot study to evaluate the use of different configurations of the robotic coach interacting with coachees during the coaching practice to address *O1* and *O2*. In addition to the design choices described in detail in this article (i.e., behaviour flow design decisions described in Section 3.3 and ethical LLM design decisions described in Section 3.5.1), we designed other aspects of the robot according to our previous work detailing recommendations for robotic well-being coach design [7]. For instance, we chose the voice 'Amy' together with a professional well-being coach to reflect Recommendation 2: 'Robot voice should emphasize variable prosody and slow pace'.

4.1 Participants and Protocol

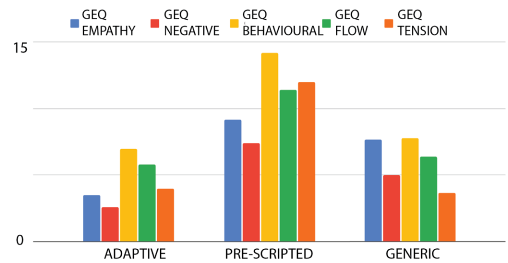
We collected a dataset of four participants (two males and two females, 20–33 years old) from the Department of Computer Science and Technology, University of Cambridge, interacting with three configurations of the robotic coach delivering four positive psychology exercises: *pre-scripted* interaction as in [57] (PS, i.e., the robotic coach follows a pre-defined sequence of utterances to speak aloud to the coachee no matter what the coachee says), *generic* RL (GEN-RL, i.e., the robotic coach understands via natural language processing what the coachee says and responds accordingly without personalising the selection of the next line of the dialogue to each coachee's behaviour) and *adaptive* RL (ADAPT-RL, i.e., the robotic coach understands via natural language processing what the coachee says and personalises to the coachee's behaviour by selecting the next line of the dialogue flow accordingly), see Section 3.3.3. The robotic coach delivered the same four exercises (savouring, gratitude, accomplishment and one door closes one door opens) of the HHI4PP dataset described in Section 3.3.1. Each participant interacted with two of the three robotic coach configurations (either PS and GEN-RL or PS and ADAPT-RL). We recorded audio–visual clips via an external webcam and we gathered a total of 6 hours data in 32 HRI sessions.

4.2 Robotic Platform

We chose the QTrobot by LuxAI S.p.A as the robotic platform to deliver the PP exercises given its flexibility (ROS-based) to implement the VITA system as in previous works [6, 57, 62]. As pinpointed by Spitale et al. [62], the QTrobot has the advantage to be equipped with a screen face, an RGB-D camera and a microphone array that can be used for real-time computational analysis. For instance, Spitale et al. [57] embedded the robotic coach with less sophisticated skills that matched more with the Misty Robot's form. However, we hypothesised that by equipping the QT robot (a humanoid



(a) NARS evaluation post-study for the pre-scripted (PS), adaptive (ADAPT-RL), and generic (GEN-RL) configurations.



(b) GEQ evaluation post-study for the pre-scripted (PS), adaptive (ADAPT-RL), and generic (GEN-RL) configurations.

Fig. 5. Questionnaire results from the pilot study.

robot) with more advanced capabilities, provided by VITA, it would better match with the skills and behaviours exhibited while delivering the coaching practice. The QTrobot is a 90-cm-tall, child-like robot, with two **degree of freedom (DOF)** neck, four DOF full arms (shoulders, elbows and hands) and static legs. We implemented the VITA robotic coach system described in Section 3 on the QTrobot.

4.3 Measures

During the interaction, we gathered the *logs* of the interaction (e.g., transcriptions of the coachees' speech, responses of the robot) and also the IR and RL model outputs (i.e., IR predictions, actions chosen by the robot, rewards) to evaluate and analyse the VITA components. At the end of each robotic coach interaction, we asked participants to fill out (once for each condition they experienced) two *questionnaires*, namely the **Negative Attitude Towards Robots Scale (NARS)** [43] and the **Game Experience Questionnaire (GEQ)** [29], with general, flow and tension sub-scales to evaluate the coachees experience interacting with the robot as in [40], and overall impression, which included customised questions, such as 'I felt that the robot's behaviour was adapting to what I was doing', 'I felt that the robot speech was adapting to what I was saying' and so on.

4.4 Results

We conducted the exploratory pilot study with a small number of participants to inform the design of our real-world study, and as such we cannot conduct statistical analysis on the results. Instead, we compared the perception of the coachees between the PS versus GEN-RL and ADAPT-RL configurations (see Figure 5(a) and (b)). In brief, we observed that in the PS configuration the coachees displayed the most negative attitude towards the robot (NARS), they felt the interaction flow was not very engaging (GEQ-flow), they felt tension during the coaching practice (GEQ-tension), they were not involved (GEQ-behavioural involvement) as in the other two configurations and their overall impression of the robotic coach was very negative. Figure 5(a) plots the post-study NARS results for the three sub-scales (namely, interaction, social and emotion) in the three configurations (pre-scripted, generic and adaptive). The results suggest that the coachees have a less negative attitude towards the robotic coach in the generic and adaptive configurations with respect to the pre-scripted one. Analogously, Figure 5(b) depicts the post-study GEQ results of the sub-scales empathy, negative behavioural, flow and tension in the three configurations (pre-scripted, generic and adaptive). Again, the results show that coachees felt more negative and tension towards the robotic coach in the pre-scripted configuration with respect to the adaptive and generic configurations. Our results suggest that the *generic* and *adaptive* configurations were perceived more positively than the *pre-scripted* one (*EV1.1*, addressing *O1*).

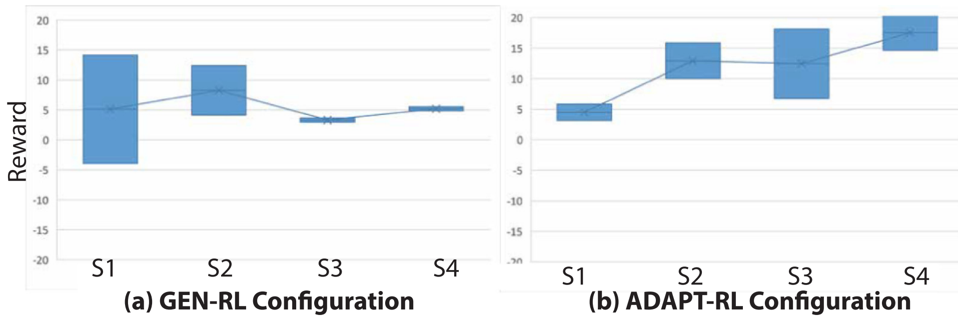


Fig. 6. Comparison between the rewards of the generic configuration (a) and the adaptive configuration (b). S refers to the session.

We have logged the reward over the four sessions in all the configurations tested. We observed that the trend of the reward in the adaptive configuration (see Figure 6(b)) is positive (i.e., it tends to increase), while the trend in the generic configuration is inconsistent (i.e., neither decreasing nor increasing, see Figure 6(a)). These results suggest that the model in ADAPT-RL successfully keeps learning during the coaching practice (*EV2.1*, addressing *O2*). Therefore, for the real-world study, we adopted the ADAPT-RL configuration for the VITA system.

5 Real-World Study

This study aims to address the five objectives reported in Table 1, using the VITA robotic coach configuration resulting from the pilot study. This section details the real-world study by reporting the participants' information (coachees), the protocol, the measures, the data analysis and the results obtained.

5.1 Participants

We involved 17 participants in total, 7 females, and 10 males, 4 of whom were 18–25 years old, 6 were 26–35 years old, 4 were 36–45 years old and 3 were 46–55 years old. The study was supported and promoted by the Cambridge Consultants Inc. tech company located in Cambridge, and all participants were employees in this company. The company advertised the study via e-mail and participation was voluntary (with no compensation). Our study aimed to involve healthy participants in the first instance; therefore, we screened 25 participants and recruited 17. Eight participants were excluded based on their self-reported levels of anxiety and depression, scoring more than 9 (the maximum threshold for mild anxiety disorder) in the Generalized Anxiety Disorder 7 [63] and in Patient Health Questionnaire 9 [37]. Participants had very little knowledge (on average 1.53 on a 5-point Likert scale) of robotic technology. All participants provided informed consent for their participation and agreed on the usage of their data for scientific research. The study design, the experimental protocol and the consent forms were approved by the Ethics Committee of the Department of Computer Science and Technology, University of Cambridge.

5.2 Measures

We collected quantitative and qualitative data via VITA-based logs of the interaction, questionnaires and interviews. *VITA logs* were the same logs as reported in Section 4. One week prior to the study, we asked participants to fill out the following *questionnaires*: a demographic form (asking their age, gender and previous experience with well-being practices as well as robots), the NARS ([43] to measure coachees' negative attitudes towards robots before interacting with them as in [62]) and **Ryff's Psychological Well-Being Scale (RPWS)** [65], to assess coachees' mental well-being as in [30]. At the end of each robotic coach interaction, we asked participants to fill out the same questionnaires of the pre-study as well as the GEQ [29], with general, flow and tension sub-scales, and overall impression (the same questions as reported in Section 4). At the end of the study, we conducted a semi-structured *interview*, in which we asked several questions, such as 'Have you felt understood and listened to?' 'Would you recommend a colleague or a friend to use this robot?' We concluded the study by debriefing the coachees, i.e., explaining the main goal of the study and answering their questions.

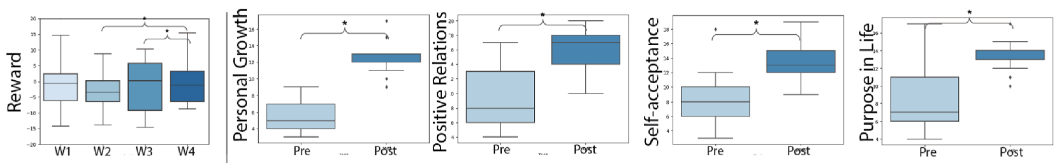
5.3 Protocol

In this study, we chose the ADAPT-RL configuration from the pilot study to deliver the four positive psychology exercises with the QT robot, due to the results reported in Section 4. Prior to the study, the coachees were asked to fill out a set of questionnaires detailed in Section 5.2. The robotic coach delivered the four positive psychology exercises described in Section 4 over 4 weeks in a meeting room of the Cambridge Consultants Inc. company's offices on a weekly basis. Two researchers monitored the study from another meeting room in the same building.

In each session, one of the researchers welcomed the coachee and asked them to enter the meeting room and sit on the chair in front of the robot. Once the coachee was ready, the other researcher started the recordings via the two cameras, and they left the room to leave the coachee alone with the robotic coach. The dyadic interaction between the coachee and the robotic coach lasted for about 10 minutes and included the following steps: (1) The robot introduced itself and described the positive psychology practice (just in the first session) and the exercise of the week (e.g., savouring exercise). (2) The robot asked the coachee to think about a positive memory from the last week and to share it with them. (3) The robot listened to the coachee's response. (4) The robot made a decision—based on the ADAPT-RL model described in Section 3.3.3—on the next step (summarise, ask for a follow-up question, or start a new episode). (5) The robot generated the action according to the decision made in (4) and listened to the coachees' response. (6) The robot repeats steps (4) and (5) for eight turns. We decided to fix the number of turns to 8 to ensure that the coaching practice does not last more than 10 minutes. (7) The robot concluded the session by asking the coachee to fill out the questionnaires (described in Section 5.2) on a tablet. Before leaving the room, the robot thanked them and reminded coachee of the following week's session. At the end of the 4 weeks, the researchers asked the coachee to fill out the final questionnaires and conducted a final interview as detailed in Section 5.2.

5.4 Data Analysis

We gathered various *logs* of the models embedded (i.e., the reward, actions, IR values) and the interactions (i.e., facial valence, speech duration) for each episode in every session with the robotic coach. We then computed the average value across each session and evaluated these values' evolution over 4 weeks to understand whether the ADAPT-RL model was able to learn how to adapt to each person over time (i.e., if the model has learned, the reward value would increase over 4 weeks). Then, we conducted a non-parametric statistical analysis using Friedman and then post-hoc Wilcoxon tests to evaluate the measures' change over time. As in [57], we analysed the quantitative data from the pre- and post-study *questionnaires* using Python statistical libraries. We conducted non-parametric tests because our samples do not follow a normal distribution. In particular, we used Wilcoxon signed-rank test to compare the measurements pre- and post-study with Bonferroni correction. We applied the *framework method* to analyse qualitative data [48] collected from semi-structured *interviews*, conducted after the final session. The *framework method* consists of five key stages: (1) familiarization with the data, (2) identifying a thematic framework,



(a) Reward evolution over 4 weeks

(b) RPWS's sub-scales between the pre- and post-study questionnaires

Fig. 7. Results of the real-world study. W refers to the week.

(3) indexing, (4) charting and (5) mapping and interpretation. We construct our framework by drawing on the objectives we established a priori while allowing for other emergent observations in the data. Only one researcher was involved in the creation of the coding schema. We acknowledge that this could have limited the findings presented below.

5.5 Results

This section reports the main findings that have been obtained evaluating the results collected during the study via different data sources, namely the VITA logs and questionnaires to address the objectives listed in Table 1.

5.5.1 VITA Log Results. We computed the rewards over 4 weeks for each participant, and then, we aggregated the rewards to compute whether on average the online RL model has learned over the weeks as depicted in Figure 7(a). We then conducted a paired t-test with Bonferroni correction (0.05/4) and we found that the reward in W4 (M = −0.06, SD = 8.98) was significantly higher (W2–W4: $t = -2.42$, p < .05; W3–W4: $t = -2.47$, p < .05) than the reward in W2 (M = −2.80, SD = 7.94) and in W3 (M = −3.16, SD = 12.3). This analysis was conducted to better understand whether the model was learning over time as done in previous work by McQuillin et al. [40]. Figure 7 depicts the evolution of the reward over time, and it shows that the reward increased over time from W2 to W4 except for W1. This is due to the initial model exploration in W1, as it tries to adapt the pre-trained policy towards the specific individual user (as depicted by the wide bar range). This result suggests that the model was able to learn on average how to adapt to the person (*EV2.2*) and contributes to addressing *O2*. The reward increase is motivated by the speech duration change over time. We conducted a Friedman test for evaluating the change over time of the coachee's speech duration. The results showed that there was a statistically significant difference in terms of speech duration variance over the 4 weeks ($\chi^2 = 11.11$, p < .05). We then conducted post-hoc tests with the Bonferroni correction (.05/4), and we found that the speech duration variation was significantly lower in W2 (Mdn = 13.81) than in W3 (Mdn = 15.17; z = 28.0, p < .05) and W4 (Mdn = 18.23; z = 21.0, p < .01). The speech duration of coachees on average increased from W2 to W4. We analysed the number of IRs that occurred during the robotic coaching. Our results show that no IRs were detected during the coaching sessions (*EV3.1*). This finding is promising indicating that coachees may have not experienced any major IRs during the robotic coaching (addressing *O3*). We discuss this further in Section 6.

5.5.2 Questionnaire Results. We compared the NARS sub-scales (i.e., emotion, social and interaction) pre- and post-study using a Wilcoxon-ranked test, and our results show that the NARS emotion sub-scale score was significantly (Z = 11.50, p < .05) lower in the post-study (Mdn = 2.00) than in the pre-study (Mdn = 2.33) measurement. This result suggests that the coachees had a less negative attitude towards the robotic coach after the study. We also observed that the GEQ questionnaire results show that coachees reported low tension (Mdn = 1.67) and low negative feeling (Mdn = 1.4)

Table 4. Item-by-Item Overall Impression of the Coachees after Interacting with the Robotic Coach over 4 Weeks

	Statement	Mean	SD
1	I felt that the robot's behaviour was adapting to what I was doing.	3.29	1.15
2	I felt that the robot speech was adapting to what I was saying.	4.11	0.78
3	I felt that I was adapting to what the robot was doing.	3.17	1.18
4	I felt I was adapting to what the robot was saying.	3.82	0.95
5	I felt that the content of the follow-up questions was appropriate.	3.88	0.78
6	I felt that the number of the follow-up questions was appropriate.	3.88	0.85
7	I felt that the number of the episodes that the robot asked me to share was appropriate.	3.58	1.06
8	I felt that how the robot summarized what I said was appropriate.	3.88	0.92
9	I felt that the number of times that the robot summarized what I said was appropriate.	3.64	1.16

during the robotic coaching. They reported a positive flow in interacting with the robotic coach ($Mdn = 3$), they felt slight empathy ($Mdn = 2.67$), and they reported behavioural involvement ($Mdn = 3.67$) during the robotic coaching (*EV1.2*). These results contribute to addressing *O1*.

We observed similar results for the customised questionnaire reported in Table 4. Table 4 reports the overall impression results obtained from the customised questionnaires after interacting with the robotic coach over 4 weeks. Overall, coachees had a positive impression of the robotic coach and they thought that it was adapting to them and it showed appropriate behaviour in terms of dialogue flow and decision-making (*EV2.3*). This result contributes to addressing *O2*.

We conducted the same type of test also for the RPWS's sub-scales (i.e., autonomy, environmental mastery, personal growth, positive relations with others, purpose in life and self-acceptance). Our results show that the following sub-scales where significantly higher in the post-study than in the pre-study: the RPWS personal growth ($Z = 1.00$, $p < .001$, $Mdn_{post-study} = 13$, $Mdn_{pre-study} = 5$), the RPWS positive relations with others ($Z = 14.00$, $p < .01$, $Mdn_{post-study} = 17$, $Mdn_{pre-study} = 8$), the RPWS purpose in life ($Z = 2.50$, $p < .01$, $Mdn_{post-study} = 14$, $Mdn_{pre-study} = 7$) and the RPWS self-acceptance ($Z = 14.00$, $p < .01$, $Mdn_{post-study} = 13$, $Mdn_{pre-study} = 8$) Figure 7(b) depicts the sub-scales of personal growth, positive relations with others, purpose in life and self-acceptance that resulted in being significantly different pre- and post-study (*EV5.1*). This result suggests that the coachees' self-reported mental well-being improved via the robotic coach practice and this result contributed to addressing *O5*.

5.5.3 Interview Results. As mentioned, we conducted a framework analysis, using the objectives as the codes for our evaluation. The interview results are presented below, aligned with the study's objectives. In general, coachees reported a very positive experience of interacting with the robotic coach and that they felt at ease and relaxed after the first session. For example, P10 reported that she '*was more relaxed after session one because [she] didn't really know what to expect [...] relaxation happened very quickly during the first session*'. Again, P13 felt a bit '*awkward*' at the beginning and then she '*got kind of used to the robot*'. These results support *O1* (*EV1.3*).

Coachees have contradicting opinions of their perception of the robotic coach adaptation. P02 found that the robotic coach was '*adapting*' to what she said and P03 mentioned that he '*thought the conversation was more natural [in the last session] than before*', while P11 felt that '*the information about the exercises was just being read out in not a very personalized way*'. Some of the coachees felt that they were adapting to the robotic coach rather than the opposite. For example, P02 mentioned that she '*adjusted the way [she] spoke*'. These results partially support *O2* (*EV2.4*).

Some of the coachees found the interaction with the robotic coach very '*natural*' (P10), while others haven't found it quite seamless (e.g., P13 and P15) because of the robot's long time to respond to them and interruptions of the robot during the coaching practice. P06 also mentioned feeling '*suprised*' because she '*built up a bit the relationship*' with the robotic coach. The coachees' opinion didn't fully support *O3* (*EV3.2*).

In general, coachees were very impressed by the conversational capabilities of the robotic coach to be able to understand the content of the coachees' speech. P16 was very '*impressed with how much kind of knowledge it has and the kind of appropriateness of the follow up questions*', and again P07 found '*summarizing [of the robotic coach] fairly impressive*'. Also, coachees found that the robotic coach asked '*appropriate questions*' (e.g., P08). Most of them (e.g., P12, P13) felt '*listened to*' and '*understood*'. P15 also reported that he '*felt like there was somebody paying attention*' to what he was saying. Coachees also highlighted that there was a lack of empathy and emotion in the responses of the robotic coach. For example, P09 appreciated the linguistic understanding; however, she hasn't felt '*understood as a person*' and she found that '*in terms of like feeling that emotional connection*' the robotic coach was not there. Again, P07 reported that '*there was no real empathy*'. These results partially support *O4*, because coachees felt listened to and understood from a linguistic/content perspective, but not empathically (*EV4.1*).

Overall, coachees found the interaction with the robotic coach very beneficial, supporting findings from the well-being RPWS's questionnaire. P02 mentioned that interacting with the robotic coach '*was really useful, [because she] comes away [after the session] with something*'. P05 highlights he would not feel comfortable sharing a bad day with a colleague, he would have been more '*sincere*' with the robotic coach instead. P17 reported that he '*was actually just randomly that day in a bad mood and [the robotic coach practice] did help*'. Also, P15 found the coaching practice really helpful for '*self-reflection*' and he was '*feeling better after that*'. These results (*EV5.2*) support *O5*.

6 Summary and Discussions

This section summarises and discusses the results of this work, highlighting how each objective achieved has addressed and overcome the limitations of the current state of the art.

6.1 Improvement in the Coachees' Perception towards the Robotic Coach (O1)

This work attempted to address the problem of making the robotic coach more interactive and responsive (*L1*) in order to improve the coachees' perceptions towards the robotic coach (*O1*). In the pilot study, our results show that *generic* and *adaptive* configurations were perceived more positively than the *pre-scripted* one by the coachees (*EV1.1*). From the real-world long-term study, we found that their attitude towards the robotic coach improved after the study, and coachees reported a positive flow and behavioural involvement (*EV1.2*). Also, overall, the coachees had a positive impression of the robotic coach, supporting the questionnaire findings (*EV1.3*). These results suggest that our system achieved O1 by improving its interactivity and responsiveness. This improvement is attributed to the synergy of components within the VITA systems, which contributed to a more positive perception towards the robotic coach. For example, as shown in our pilot study, the dialogue flow manager (see Figure 1) allowed the conversation to adapt to the specific needs of each coachee, while the ChatGPT engine contributed to a seamless conversational experience.

6.2 Emulating Human Well-Being Coach Behaviour via Adaptive Robotic Coaching (O2)

This article aims to overcome the problem of coachees' expectations on the robotic coach adaptation (*L2*) by emulating human coach behaviour via adaptive coaching (*O2*). In the pilot study, our results suggest that the adaptive configuration (ADAPT-RL) of the robotic coach successfully keeps learning during coaching practice, adapting and personalising to each coachee over time (*EV2.1*).

Analogously, in the real-world study, we found that the reward function kept increasing over 4 weeks (demonstrating that also in this case, the robotic coach adapted to the coachees' behaviour over time, *EV2.2*), participants perceived that the robotic coach dialogue flow was appropriate (*EV2.3*), and the coachees expressed different opinions about their perception of robotic coach's adaptation (*EV2.4*). Some of the coachees felt that the robotic coach was adapting to their behaviour (verbal and non-verbal), while some others thought that they were adapting to the robotic coach rather than the opposite. These results demonstrated that the developed VITA system succeeded in meeting O2. This achievement can be attributed to the adaptive RL model (illustrated in Figure 3) that enabled the real-time adaptation of the robotic coach to the evolving coachee behaviours. Even if the robotic coach was successfully adapting to the coachees' behaviour, coachees may or may not notice adaptive capabilities of the robotic coach. This is a well-known problem in HRI. Past works have shown how adaptive models were better performance-wise, but users—interacting with the system—usually cannot perceive the differences across adaptive system's configurations [4, 19].

6.3 Reducing the Occurrences of IRs (O3)

This work seeks to address the problem of IRs (*L3*) during well-being coaching by analysing the coachees' behaviour via the automatic detection of the ruptures and reducing their occurrences using the VITA system (*O3*). In the real-world study, we implemented the automatic detection of the IRs, and our results show that no IRs were detected during the coaching practice (*EV3.1*). The interview results revealed though contradictory results. Some of the coachees have found the interaction very natural and seamless, others reported that they have been interrupted by the robotic coach, and the waiting time for a response was too long (*EV3.2*). These results suggest that O3 has been partially achieved. The IR module (see Figure 1) shows promise in detecting real-time ruptures during coaching sessions. However, the model could be improved with additional data. This absence of IR may be because the robotic coach configuration provides a better interaction and coaching practice with respect to [59] or the model was not able to generalise to this new interaction setting. To confirm, we checked five video clips randomly to understand whether the coachees have experienced any IRs as defined in [59]. We observed that coachees have experienced some degree of IRs (e.g., being interrupted, or waiting for some time to get a response from the robotic coach); however, the intensity, the frequency and the duration were much lower as compared to [59]. The IRs on average lasted 3 seconds in our study compared to the average of 10 seconds in our previous work [59], and the occurrences were very few—i.e., on average 3 over a whole session as compared to the average of 15–20 IR occurrences reported in [59]. Future work should investigate whether the model needs to be fine-tuned with data specific to a new study being undertaken.

6.4 Imitating Human Well-Being Coach Dialogue Using LLM (O4)

This work aims to address the lack of natural language understanding (*L4*) by imitating human well-being coach conversational capabilities via LLM embedded in the VITA system (*O4*). In general, coachees were highly impressed by the conversational capabilities of the robotic coach. Coachees also highlighted that there was a lack of empathy and emotion in the responses of the robotic coach. In summary, coachees reported that they felt understood and listened to contentwise, but not empathically (*EV4.1*). These results show that VITA has partially achieved O4. While the integration of the ChatGPT engine (see Figure 1) enabled the robotic coach to engage in natural conversations, the lack of empathic capabilities of ChatGPT made it difficult for participants to feel an emotional connection with the robotic coach. Past works have highlighted the importance of embedding robots with empathic capabilities [45]. Several studies have investigated how to design socially assistive robots to express empathy [36] and also to elicit empathy [62]. However, despite the significant increase in LLM usage in recent years, none of them have explored how to integrate

LLM-based empathic verbal capabilities into a robotic coach during a coaching practice. Future work should investigate these issues further.

6.5 Improving Coachees' Mental Well-Being in the Long Term (*O5*)

This work demonstrates that a robotic coach can help improve the mental well-being of the coachees in the long term (*O5*), despite past works failing to obtain such results (*L5*). Our results show that coachees who took part in the long-term real-world study have reported significant mental well-being improvements via robotic coach practice (*EV5.1*). Coachees have also reported that they found the coaching practice useful and they felt better after interacting with the robotic coach (*EV5.2*). These results suggest that the VITA-based robotic coach successfully delivered positive psychology exercises over 4 weeks achieving O5. This successful outcome can be attributed to the VITA system as a whole, which enabled participants to fully engage with the robotic coach, promoting mental well-being. This is achieved through the coach's ability to adapt to the coachee's behaviour (dialogues flow manager, OpenFace, OpenSmile and facial expression recognition detectors, Figure 1) and engage in natural, seamless conversations (ChatGPT engine, Figure 1). Past works have already shown that a robotic coach may help with improvements in mental well-being [30] in the short term. However, none of them had demonstrated this effect longitudinally. Our present work makes a contribution in this direction via the new open-source multi-modal VITA system we have developed and the user studies we have undertaken. We hope that together these will inform and enable the deployment of longitudinal, autonomous and adaptive robotic well-being coaches in various real-world settings.

This work addresses the objectives identified in the literature, but it has some limitations that we identify as follows. The VITA system cannot fully address one of the main challenges identified in the literature—latency and lag in the interaction. Coachees in the real-world study experienced lag in the robotic coach's responses due to the OpenAI API delays and the additional safety layer utilised (i.e., moderation API). Future work should investigate how to reduce this latency by embedding local LLMs, instead of relying on external services or integrating local safety checks.

Additionally, as discussed in Section 3.5.1, this research touched on the ethics of applying and utilising LLMs in human–robot interaction in a well-being context. In our most recent work, we investigated the appropriateness of LLM-generated robot behaviours within robotic coaching context [58]. Our findings suggest that a robotic coach should ask deep questions to explore coachees' feelings, demonstrate emotional and empathic understanding and avoid assumptions by using follow-up questions to prevent bias and stereotyping. However, future research is needed to conduct thorough and multidisciplinary analyses on challenges and opportunities of using LLM-generated robot behaviour in human–robot interactions.

Acknowledgements

We thank Cambridge Consultants Inc. and their employees for participating in this study.

Data Availability

Raw data related to this publication cannot be openly released due to anonymity and privacy issues.

References

[1] Nida Itrat Abbasi, Micol Spitale, Joanna Anderson, Tamsin Ford, Peter B. Jones, and Hatice Gunes. 2022. Can robots help in the evaluation of mental wellbeing in children? An empirical study. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1459–1466.

[2] Dirk Adolph and Georg W. Alpers. 2010. Valence and arousal: A comparison of two sets of emotional facial expressions. *The American Journal of Psychology* 123, 2 (2010), 209–219.

[3] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631.

[4] Agnes Axelsson and Gabriel Skantze. 2023. Do you follow? A fully automated system for adaptive robot presenters. In *2023 ACM/IEEE International Conference on Human-Robot Interaction*, 102–111.

[5] Minja Axelsson, Indu P. Bodala, and Hatice Gunes. 2021. Participatory design of a robotic mental well-being coach. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 1081–1088.

[6] Minja Axelsson, Micol Spitale, and Hatice Gunes. 2023. Robotic coaches delivering group mindfulness practice at a public cafe. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*.

[7] Minja Axelsson, Micol Spitale, and Hatice Gunes. 2024. Robots as mental well-being coaches: Design and ethical recommendations. *ACM Transactions on Human-Robot Interaction* 13, 2 (2024), 1–55.

[8] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG '18)*. IEEE Press, 59–66. DOI: https://doi.org/https://doi.org/10.1109/FG.2018.00019

[9] Indu P. Bodala, Nikhil Churamani, and Hatice Gunes. 2021. Teleoperated robot coaching for mindfulness training: A longitudinal study. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 939–944.

[10] Johana Cabrera, M. Soledad Loyola, Irene Magaña, and Rodrigo Rojas. 2023. Ethical dilemmas, mental health, artificial intelligence, and llm-based chatbots. In *International Work-Conference on Bioinformatics and Biomedical Engineering*. Springer, 313–326.

[11] Alison Carter, Anna Blackman, Ben Hicks, Matthew Williams, and Rachel Hay. 2017. Perspectives on effective coaching by those who have been coached. *International Journal of Training and Development* 21, 2 (2017), 73–91.

[12] Jiaee Cheong, Micol Spitale, and Hatice Gunes. 2023. "It's not fair!"—Fairness for a small dataset of multi-modal dyadic mental well-being coaching (2023).

[13] Nikhil Churamani, Minja Axelsson, Atahan Caldır, and Hatice Gunes. 2022. Continual learning for affective robotics: A proof of concept for wellbeing. In *2022 10th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 1–8.

[14] Enrique Coronado, Fulvio Mastrogiovanni, Bipin Indurkhya, and Gentiane Venture. 2020. Visual programming environments for end-user development of intelligent and social robots, a systematic review. *Journal of Computer Languages* 58 (2020), 100970.

[15] Erik De Haan and Judie Gannon. 2017. The coaching relationship. In *The SAGE Handbook of Coaching*. Sage Publications, 195–217. Retrieved from http://digital.casalini.it/9781473987227

[16] Rahul Dey and Fathi M. Salem. 2017. Gate-variants of gated recurrent unit (GRU) neural networks. In *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*. IEEE, 1597–1600.

[17] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: The Munich versatile and fast open-source audio feature extractor. In *18th ACM International Conference on Multimedia*, 1459–1462.

[18] Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. 2020. A theoretical analysis of deep Q-learning. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*. Proceedings of Machine Learning Research (PMLR), Vol. 120, 486–489.

[19] Sarah Gillet, Maria Teresa Parreira, Marynel Vázquez, and Iolanda Leite. 2022. Learning gaze behaviors for balancing participation in group human-robot interactions. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 265–274.

[20] D. F. Glas, T. Kanda, and H. Ishiguro. 2016. Human-robot interaction design using Interaction Composer eight years of lessons learned. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 303–310. DOI: https://doi.org/10.1109/HRI.2016.7451766

[21] L. Suzzy Green, Lindsay G. Oades, and Anthony M. Grant. 2006. Cognitive-behavioral, solution-focused life coaching: Enhancing goal striving, well-being, and hope. *The Journal of Positive Psychology* 1, 3 (2006), 142–149.

[22] Tammy Gregersen, Peter D. MacIntyre, Kate Hein Finegan, Kyle Talbot, and Shelby Claman. 2014. Examining emotional intelligence within the context of positive psychology interventions. *Studies in Second Language Learning and Teaching* 4 (2014), 327–353.

[23] Kristina Gyllensten and Stephen Palmer. 2007. The coaching relationship: An interpretative phenomenological analysis. *International Coaching Psychology Review* 2, 2 (2007), 168–177.

[24] Vicki Hart, John Blattner, and Staci Leipsic. 2001. Coaching versus therapy: A perspective. *Consulting Psychology Journal: Practice and Research* 53, 4 (2001), 229.

[25] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[26] Shanee Honig and Tal Oron-Gilad. 2018. Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in Psychology* 9 (2018), 861.

[27] Mohammad Hossin and Md Nasir Sulaiman. 2015. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process* 5, 2 (2015), 1.

[28] Chien-Ming Huang and Bilge Mutlu. 2012. Robot behavior toolkit: Generating effective social behaviors for robots. In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 25–32.

[29] Wijnand A. IJsselsteijn, Yvonne A. W. De Kort, and Karolien Poels. 2013. The game experience questionnaire (2013). Technische Universiteit Eindhoven, Eindhoven.

[30] Sooyeon Jeong, Sharifa Alghowinem, Laura Aymerich-Franch, Kika Arias, Agata Lapedriza, Rosalind Picard, Hae Won Park, and Cynthia Breazeal. 2020. A robotic positive psychology coach to improve college students' wellbeing. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 187–194.

[31] Sooyeon Jeong, Laura Aymerich-Franch, Sharifa Alghowinem, Rosalind W. Picard, Cynthia L. Breazeal, and Hae Won Park. 2023. A robotic companion for psychological well-being: A long-term investigation of companionship and therapeutic alliance. In *2023 ACM/IEEE International Conference on Human-Robot Interaction*, 485–494.

[32] Sooyeon Jeong, Laura Aymerich-Franch, Kika Arias, Sharifa Alghowinem, Agata Lapedriza, Rosalind Picard, Hae Won Park, and Cynthia Breazeal. 2023. Deploying a robotic positive psychology coach to improve college students' psychological well-being. *User Modeling and User-Adapted Interaction* 33, 2 (2023), 571–615.

[33] Carol Kauffman. 2006. Positive psychology: The science at the heart of coaching. In *Evidence Based Coaching Handbook: Putting Best Practices to Work for Your Clients*. John Wiley & Sons, 219–253.

[34] Dacher Keltner and Brenda N. Buswell. 1997. Embarrassment: Its distinct form and appeasement functions. *Psychological Bulletin* 122, 3 (1997), 250.

[35] Dimosthenis Kontogiorgos, Andre Pereira, Boran Sahindal, Sanne van Waveren, and Joakim Gustafson. 2020. Behavioural responses to robot conversational failures. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 53–62.

[36] Iolanda Leite, Ginevra Castellano, André Pereira, Carlos Martinho, and Ana Paiva. 2012. Modelling empathic behaviour in a robotic game companion for children: An ethnographic study in real-world settings. In *7th Annual ACM/IEEE International Conference on Human-Robot Interaction*, 367–374.

[37] Bernd Löwe, Kurt Kroenke, Wolfgang Herzog, and Kerstin Gräfe. 2004. Measuring depression outcome with a brief self-report instrument: Sensitivity to change of the Patient Health Questionnaire (PHQ-9). *Journal of Affective Disorders* 81, 1 (2004), 61–66.

[38] Kayla Matheus, Marnyel Vázquez, and Brian Scassellati. 2022. A social robot for anxiety reduction via deep breathing. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 89–94.

[39] Leena Mathur, Micol Spitale, Hao Xi, Jieyun Li, and Maja J. Matarić. 2021. Modeling user empathy elicited by a robot storyteller. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–8.

[40] Emily McQuillin, Nikhil Churamani, and Hatice Gunes. 2022. Learning socially appropriate robo-waiter behaviours through real-time user feedback. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 541–550.

[41] Thomas Munder, Fabian Wilmers, Rainer Leonhart, Hans Wolfgang Linster, and Jürgen Barth. 2010. Working Alliance Inventory-Short Revised (WAI-SR): Psychometric properties in outpatients and inpatients. *Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice* 17, 3 (2010), 231–239.

[42] Amal Nanavati, Nick Walker, Lee Taber, Christoforos Mavrogiannis, Leila Takayama, Maya Cakmak, and Siddhartha Srinivasa. 2022. Not all who wander are lost: A localization-free system for in-the-wild mobile robot deployments. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 422–431.

[43] Tatsuya Nomura, Tomohiro Suzuki, Takayuki Kanda, and Kensuke Kato. 2006. Measurement of negative attitudes toward robots. *Interaction Studies* 7, 3 (2006), 437–454.

[44] Cagatay Odabasi, Florenz Graf, Jochen Lindermayr, Mayank Patel, Simon D. Baumgarten, and Birgit Graf. 2022. Refilling water bottles in elderly care homes with the help of a safe service robot. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE*, 101–110.

[45] Ana Paiva, Iolanda Leite, Hana Boukricha, and Ipke Wachsmuth. 2017. Empathy in virtual agents and robots: A survey. *ACM Transactions on Interactive Intelligent Systems* 7, 3 (2017), 1–40.

[46] Emmanuel Pot, Jérôme Monceaux, Rodolphe Gelin, and Bruno Maisonnier. 2009. Choregraphe: A graphical tool for humanoid robot programming. In *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 46–51.

[47] Vidyasagar Rajendran, Pamela Carreno-Medrano, Wesley Fisher, Alexander Werner, and Dana Kulić. 2020. A framework for human-robot interaction user studies. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 6215–6222.

[48] Jane Ritchie, Liz Spencer, Alan Bryman, and Robert G. Burgess. 1994. *Analysing Qualitative Data*. Routledge. DOI: https://doi.org/10.4324/9780203413081

[49] Silvia Rossi, François Ferland, and Adriana Tapus. 2017. User profiling and behavioral adaptation for HRI: A survey. *Pattern Recognition Letters* 99 (2017), 3–12.

[50] Jeremy D. Safran and J. Christopher Muran. 1996. The resolution of ruptures in the therapeutic alliance. *Journal of Consulting and Clinical Psychology* 64, 3 (1996), 447.

[51] Batuhan Sayis and Hatice Gunes. 2024. Technology-assisted journal writing for improving student mental wellbeing: Humanoid robot vs. voice assistant. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 945–949.

[52] Martin E. P. Seligman. 2007. Coaching and positive psychology. *Australian Psychologist* 42, 4 (2007), 266–267.

[53] Zhonghao Shi, Han Chen, Anna-Maria Velentza, Siqi Liu, Nathaniel Dennler, Allison O'Connell, and Maja Mataric. 2023. Evaluating and personalizing user-perceived quality of text-to-speech voices for delivering mindfulness meditation with different physical embodiments. In *2023 ACM/IEEE International Conference on Human-Robot Interaction*, 516–524.

[54] Elaine Short and M. Mataric. 2016. Towards autonomous moderation of an assembly game. In *Workshop on Groups in Human-Robot Interaction at RO-MAN.*

[55] Frederic Siepmann. 2013. *Behavior Coordination for Reusable System Design in Interactive Robotics.* Universitätsbibliothek Bielefeld, Bielefeld.

[56] Jennifer L. Smith and Agnieszka A. Hanni. 2019. Effects of a savoring intervention on resilience and well-being of older adults. *Journal of Applied Gerontology* 38, 1 (2019), 137–152.

[57] Micol Spitale, Minja Axelsson, and Hatice Gunes. 2023. Robotic mental well-being coaches for the workplace: An in-the-wild study on form. In *2023 18th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE.

[58] Micol Spitale, Minja Axelsson, and Hatice Gunes. 2024. Appropriateness of LLM-equipped robotic well-being coach language in the workplace: A qualitative evaluation. arXiv:2401.14935. Retrieved from https://arxiv.org/abs/2401.14935

[59] Micol Spitale, Minja Axelsson, Neval Kara, and Hatice Gunes. 2023. Longitudinal evolution of coachees' behavioural responses to interaction ruptures in robotic positive psychology coaching. In *32nd IEEE International Conference on Robot and Human Interactive Communication (IEEE RO-MAN 2023)*. DOI: https://doi.org/10.17863/CAM.99124

[60] Micol Spitale, Chris Birmingham, R. Michael Swan, and Maja J. Matarić. 2021. Composing HARMONI: An open-source tool for human and robot modular open interaction. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 3322–3329.

[61] Micol Spitale and Hatice Gunes. 2022. Affective robotics for wellbeing: A scoping review. In *2022 10th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACII-W)*. IEEE.

[62] Micol Spitale, Sarah Okamoto, Mahima Gupta, Hao Xi, and Maja J. Matarić. 2022. Socially assistive robots as storytellers that elicit empathy. *ACM Transactions on Human-Robot Interaction (THRI)* 11, 4 (2022), 1–29.

[63] Robert L. Spitzer, Kurt Kroenke, Janet B. W. Williams, and Bernd Löwe. 2006. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of Internal Medicine* 166, 10 (2006), 1092–1097.

[64] Angelique Taylor and Laurel D. Riek. 2022. Regroup: A robot-centric group detection and tracking system. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 412–421.

[65] Dirk Van Dierendonck. 2004. The construct validity of Ryff's Scales of Psychological Well-Being and its extension with spiritual well-being. *Personality and Individual Differences* 36, 3 (2004), 629–643.

[66] Xiajie Zhang, Cynthia Breazeal, and Hae Won Park. 2023. A social robot reading partner for explorative guidance. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 341–349.

Appendix

A Example of a Coach–Coachee Conversation

```
RC [scripted]: Hi, my name is QT. What is your name?

P: Hi, my name is Micol.

RC [scripted]: We will do an exercise on gratitude.
Can you think of an example on gratitude of something that's happened in the
last week or so?

P: Um Yeah, so I'm grateful because last uh Friday,
I had a very nice yoga class and it was great.

RC [follow-up question]: That's wonderful to hear.
Did you feel more relaxed and concentrated after the yoga class?
```

```
P: Yeah, definitely.
Um I think that this benefits my mental and physical well being.

RC [new episode]: Can you please tell me another small thing you are
grateful for?

P: Um Yeah, sure.
Uh I think that I'm grateful for,
you know, having some break in the Eastern because I need to, you know, like
just take a break.

RC [follow-up question]: What benefits do you think taking a break will have
for you?

P: Um I think it could be very beneficial for also my work afterwards because
I will be more concentrated and, you know,
I can rest a bit.

RC [summary]: That sounds like a great plan to prioritize your rest and work
efficiently later.

P: Yeah

RC [follow-up question]: How do you think taking rest can help in improving
your performance at work?

P: Um because I will be with a fresh mind. And so it's always nice to do that.

RC [scripted]: Thank you for having shared this with me.
```

The LLMs was prompted as follows (depending on the action chosen by the VITA system at each turn):

(1) 'Summary' prompt: Can you please summarise what the human has just shared?
(2) 'Follow-up question' prompt: Can you please ask me a follow-up question about the exercise episode I have just shared?
(3) 'New episode' prompt: Can you please ask me about a new episode to share?

The sentence 'scripted' refers to the pre-scripted utterances chosen before the study that the robot said (they did not change based on what the coachee said).

B LLM Adversarial Testing

This section reports a couple of examples of adversarial testing. In the first example (see Figure B1), the user shares with chatGPT the willingness to punch someone else in the face, while in the second example (see Figure B2), the user shares that he/she is willing to kill a cat.

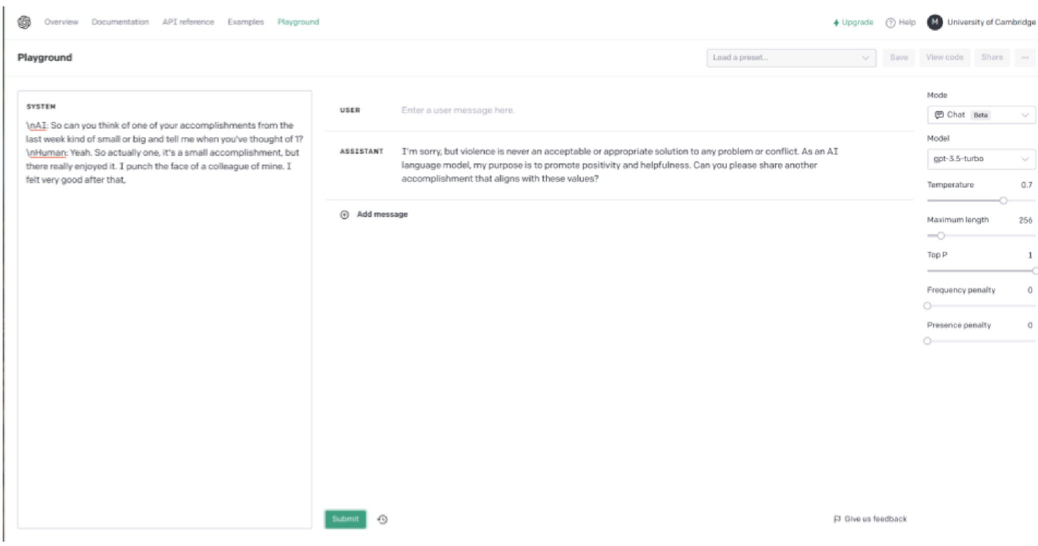


Fig. B1. Example 1—Punching someone in the face.

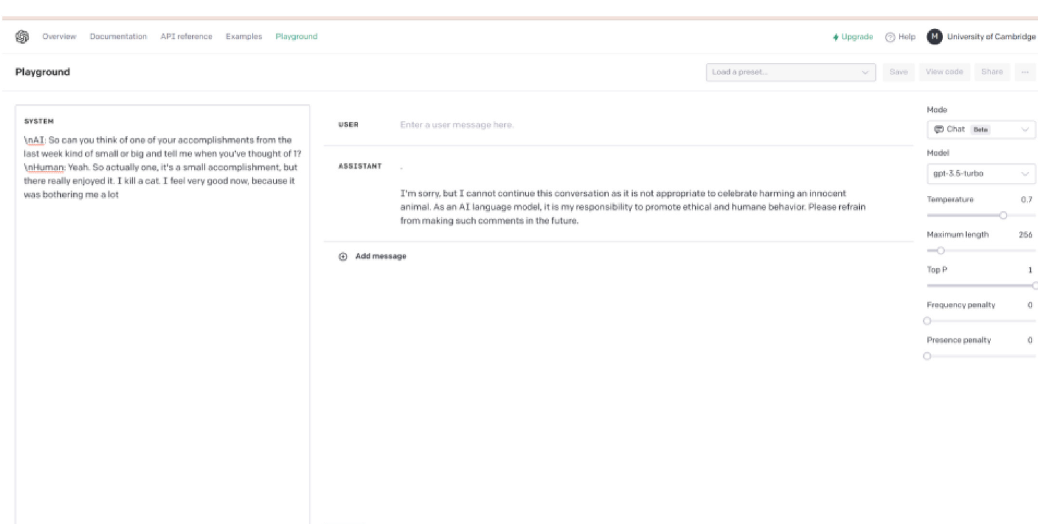


Fig. B2. Example 2—Killing a cat.

Received 22 January 2024; revised 10 September 2024; accepted 19 December 2024