

Coaching Copilot: Blended Form of an LLM-Powered Chatbot and a Human Coach to Effectively Support Self-Reflection for Leadership Growth

Riku Arakawa*
Carnegie Mellon University
Pittsburgh, USA
arakawa@cs.cmu.edu

Hiromu Yakura*
Max-Planck Institute for Human Development
Berlin, Germany
yakura@mpib-berlin.mpg.de

ABSTRACT

Chatbots' role in fostering self-reflection is now widely recognized, especially in inducing users' behavior change. While the benefits of 24/7 availability, scalability, and consistent responses have been demonstrated in contexts such as healthcare and tutoring to help one form a new habit, their utilization in coaching necessitating deeper introspective dialogue to induce leadership growth remains unexplored. This paper explores the potential of such a chatbot powered by recent Large Language Models (LLMs) in collaboration with professional coaches in the field of executive coaching. Through a design workshop with them and two weeks of user study involving ten coach-client pairs, we explored the feasibility and nuances of integrating chatbots to complement human coaches. Our findings highlight the benefits of chatbots' ubiquity and reasoning capabilities enabled by LLMs while identifying their limitations and design necessities for effective collaboration between human coaches and chatbots. By doing so, this work contributes to the foundation for augmenting one's self-reflective process with prevalent conversational agents through the human-in-the-loop approach.

CCS CONCEPTS

• **Human-centered computing** → **Interaction design; Collaborative and social computing.**

KEYWORDS

coaching, reflection, chatbot, human-AI collaboration

ACM Reference Format:

Riku Arakawa and Hiromu Yakura. 2024. Coaching Copilot: Blended Form of an LLM-Powered Chatbot and a Human Coach to Effectively Support Self-Reflection for Leadership Growth. In *ACM Conversational User Interfaces 2024 (CUI '24)*, July 8–10, 2024, Luxembourg, Luxembourg. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3640794.3665549>

1 INTRODUCTION

Self-reflection is an indispensable element for personal growth and behavior change, and research to foster it is actively conducted in

the Human-Computer Interaction (HCI) field [1, 8, 11]. In this field, chatbot-based coaching has been attracting attention with a potential to enable 24/7 support for reflection, the effectiveness of which has been demonstrated in areas like healthcare [34, 43], well-being support [31], and education [41]. While common in forming a new habit (e.g., healthy diet, exercise), using chatbots for supporting the achievement of professional goals, such as leadership growth in an organization, is less covered in prior research. Notably, such goal achievement demands a nuanced approach to fostering deep, introspective analysis of one's own behaviors and performing strategic decision-making. This paper explores the effective use of chatbots to facilitate such deep introspective reflection to augment the potential of today's conversational agents, especially in the paradigm of Large Language Models (LLMs).

In practice, *executive coaching* has played a significant role in talent development for many years [21, 64]. Executive coaching typically consists of a dialogue between a coach and their client, where the coach attempts to bring out the leadership qualities of a client through dialogues, and often involves goal setting, action planning, and accountability for professionals in high-stakes, high-stress roles [32]. The unique challenges lie in navigating complex interpersonal dynamics, understanding the subtleties of organizational cultures, and facilitating transformative insights that can guide high-level strategic decisions while maintaining clients' motivation to change behaviors [29]. As such, in recent years, there has been increasing attention in both HCI [4, 5] and human resource development (HRD) [12, 26, 62] on research to realize technological support for executive coaching.

We took executive coaching as a field to explore how LLM-powered chatbots can effectively foster one's reflection to achieve professional goals. This was motivated by prior work in HRD [26], which pointed out that merely introducing a chatbot in executive coaching encounters difficulties in identifying clients' core problems and providing individual, precise feedback. The work suggests the need to design the use of chatbots carefully with domain experts. Therefore, in this paper, we conducted two studies in collaboration with coaching organizations. First, we held a workshop with eight professional coaches, where they were asked to interact with a GPT-4-based chatbot prototype while imagining using it for coaching processes. Through the discussion, we found that the technology would be best suited to offering complementary text coaching between face-to-face sessions. We then developed an LLM-powered chatbot dedicated to such a blended experience. We conducted a two-week user study with ten pairs of coaches and clients in actual coaching situations. The semi-structured interviews with both

*These authors contributed equally and are ordered alphabetically.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

CUI '24, July 8–10, 2024, Luxembourg, Luxembourg

© 2024 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0511-3/24/07.

<https://doi.org/10.1145/3640794.3665549>

coaches and clients after the trial shed light on the advantages and limitations of chatbot coaches, as well as the importance of securing commitment with human coaches. The study supplementarily showed that the clients' behavioral intention was sustained, and their authenticity scale was affected throughout the trial. These results confirmed the benefit of conversational agents in complementing existing coaching practices for leadership growth.

Based on the study's implications, we summarized a guideline to support one's achievement of their professional goals by blending an LLM-powered chatbot coach and a human coach. While prior work suggested the necessity of the collaboration between human and artificial intelligence (AI) in chatbot-based reflection support [26], actual collaboration practice has not been well-established, particularly in such a blended manner. Our work offers a set of key factors with empirical evidence guided by both coaches and their clients in actual executive coaching sessions. Furthermore, these findings distinctly delineate the current performance boundaries of LLM-powered chatbots in fostering deep self-reflection, which demands a high level of human context understanding, and hint at the future directions of human-centered natural language processing and conversation analysis research.

2 BACKGROUND

Our research intersects with various aspects of both HRD practice and HCI research. To begin with, we offer foundational background information to set the context. We then review prior work on introducing a conversational agent to support different kinds of reflection to situate our work.

2.1 Executive Coaching and Technological Support

Executive coaching is a professional development process in HRD designed to enhance the performance of individuals in executive or leadership positions within organizations [21, 64]. This is achieved through a form of dialogue where a coach works with executives or managers in a business to help them improve their leadership skills, gain self-awareness, clarify their goals, and achieve their development objectives [32]. To this aim, typical coaching heavily relies on deep, spontaneous reflection regarding one's behaviors, beliefs, and values, leading to a clearer understanding of one's own motivations, strengths, and areas for improvement [44]. A characteristic point of executive coaching is that the dialogue is not only oriented toward immediate problem-solving or providing feedback but focused on guiding the client's self-reflection through questions [16]. Therefore, the coach's skills are essential for successful outcomes by fostering clients' readiness [25], facilitating their reflection, and maintaining their motivation [40]. While getting soaring attention, the industry faces the problem of lacking proficient coaches, preventing the practice from being affordable for everyone who needs it [28].

To make such coaching processes scalable, several technological supports have been proposed in HCI research, exploring the potential of computers to assist the process from the perspective of human-AI interaction [2]. For example, *REsCUE* [4] is a real-time feedback system for a coach during coaching sessions based on clients' multimodal behavior signals such as gaze. It was suggested

that such feedback could deepen the ongoing conversation quality from the coach's perspective. *INWARD* [5] is a video-reflection supporting system for both coach and client, successfully making the reflection process more efficient and effective for them. These systems were designed to support difficult or tedious processes within executive coaching (e.g., watching an entire video of the coaching session) while preserving the form of the existing coach-client communication. To further broaden coaching reach and cater to a larger audience in need, this study explores how intelligent conversational agents powered by recent advancements in LLMs can augment the coach-client communication.

2.2 Reflection Support in HCI

Meanwhile, reflection is a long-studied, multi-faceted concept given its importance in personal growth and behavior change [8, 11]. HCI research has tried to support users' reflection with different types of artifacts (e.g., virtual reality application [60], avatar coach [69], etc.). Examples span a wide range of contexts such as personal informatics [30], health and well-being support [31, 43, 52], parent-child interaction [33], school education [41], professional learning [66], video meeting [3, 10, 54], and creativity support [23]. For example, Mai *et al.* [41] examined the efficacy of chatbot in helping students reflect on their exam anxiety and found the benefit of computer-driven opportunity in reducing the hurdle for students to initiate reflection. Wolfbauer *et al.* [66] presented that a guidance system to support reflective writing can enhance a user's reflection competence. These various reflection-support systems were recently reviewed by Bentvelzen *et al.* [11], who concluded that the level of reflection is diverse in different contexts, calling for more empirical evidence with a situated evaluation of the reflection effect. Our research aims to contribute to the ongoing discussions by providing an example where we investigate the design and effectiveness of reflection-support technology with a strong focus on leadership growth, which requires unique, deep introspection. Considering its dialogic essence to foster reflection in contrast to methods like reflective writing via fixed questions [46], executive coaching would be particularly well-suited to explore the capabilities of LLM-powered chatbots.

2.3 Chatbot-Based Coaching

Dialogue can play a key role in our reflection process [11], as Mols *et al.* [45] found that most people conduct reflection while conversing with others. Thus, interactive chatbots have been studied as a medium to facilitate one's reflection [22], aiming to be a virtual coach in multiple areas. Specifically, the possibility has been widely examined in healthcare [9, 15, 42, 43, 50–52, 65], well-being support [14, 24, 47, 68], education [19, 36, 41, 58], and team collaboration [10, 54]. For example, Mitchell *et al.* [43] provided empirical evidence that a script-based chatbot can support individuals with type 2 diabetes. They concluded that the benefit of the chatbot coach is its persistence and consistency, fostering clients' autonomy. Similarly, Essel *et al.* [19] reported the chatbot's effectiveness in undergraduate students' learning behavior and confirmed its benefit in responding swiftly to the questions the students asked. Gabrielli *et al.* [24] presented a similar result in helping adolescents gain coping skills and mental well-being. These studies demonstrate

the power of the chatbot that is always available and not affected by external factors like humans (e.g., by fatigue or unrelated emotional events) in its responses.

As discussed in Section 2.1, executive coaching has a unique practice of having reflection through dialogue, and these benefits of chatbot may not be directly applicable. For instance, recent work in HRD argued that replacing human coaches with AIs is not optimal due to its insufficient capability to deal with diverse dialogue contexts, and the technology may work better for a set of specified topics in executive coaching [26]. Terblanche and Cilliers [57] concluded that what “works” in human-to-human coaching may not necessarily be applicable in AI-to-human coaching because users of a chatbot coach may struggle to trust an algorithm. Passmore and Tee [49] mentioned that, despite chatbots’ scalability, their inability to demonstrate empathy and emotional intelligence could be critical limitations in this sensitive domain. These papers call for empirical substantiation of a better approach to incorporating AIs than trying to alternate human coaches with chatbots.

Our work is inspired by a position paper of Graßmann and Schermuly [26] that implied the potential of augmenting traditional coaching with conversational agents rather than replacing human coaches, given the above limitations. Specifically, in this work, we seek a plausible form of blending a human coach and a chatbot coach, especially in the paradigm of the rapidly developed LLM-powered chatbots [18]. To this end, we distinctly define the capacities and roles of each party and scrutinize their impact through collaborative design with domain experts. We note that such an approach of blending a human and a chatbot coach is not yet prominent. Thus, we believe that our design considerations rooted in this unique domain can also have implications for the HCI community to diversify the ways of leveraging chatbots.

3 RESEARCH QUESTIONS

This paper explored the effective use of LLM-powered chatbots to induce deep self-reflection and behavior change beyond habit formation. For this, we took executive coaching as a study field and focused on one’s leadership growth. Throughout the rest of this paper, we seek answers to the following research questions. First, we want to discern the potential areas within the stages of the executive coaching process where chatbots can play a contributive role.

RQ1: What part of executive coaching can LLM-powered chatbot get involved with?

RQ2: What do human coaches consider as the expected role of such a chatbot and their own for successful coaching?

Then, as highlighted in Section 2.2, there is a need for empirical evidence to understand how technology aids in facilitating individual reflection. Hence, we will address the next question:

RQ3: How do clients perceive the introduction of such an LLM-powered chatbot coach in their coaching processes and exploit it?

By exploring these questions, we expected that we could derive insights into supporting professional coaches with LLM-powered chatbots in connection to prior research in HCI.

4 WORKSHOP WITH PROFESSIONAL COACHES

To answer **RQ1** and **RQ2**, we conducted a design workshop with professional coaches in June 2023. The purpose was to get their perspectives on the state-of-the-art LLM-powered chatbot concerning its capability and potential role in executive coaching.

4.1 Procedure

Eight coaches (two male and six female, 40–56 years old) were invited to our workshop via a Japanese company that offers executive coaching services to its client companies.¹ An experimenter first conducted a pre-use hearing where they informally asked their thoughts about the potential of recent chatbots, such as ChatGPT, especially in leadership growth. Then, the experimenter introduced the playground of GPT-4² and explained basic concepts of using LLM-powered chatbots, such as system prompting. Here, the coaches freely edited the prompt and had a conversation with GPT-4. After familiarizing themselves with the experience, the coaches discussed the possibility and difficulty of introducing such chatbots into executive coaching. Also, the experimenter conducted semi-structured interviews with them, asking their opinions about the discussion topics and their prospect about the role and value of human coaches after introducing the chatbots. All the above processes were conducted remotely with the attendance of the experimenter and recorded under their agreement.

4.2 Findings

In the pre-use hearing, the coaches mentioned various use cases potentially made possible by chatbots. They can be mainly categorized into supporting face-to-face and text coaching sessions. Regarding the former direction, the coaches mentioned the possibility of real-time assistance:

I appreciate it if chatbots could provide us with a selection of potential questions based on transcribing and comprehending the content of the dialogue with clients. Even if the suggestions are not optimal, they could inspire us to formulate improved questions.

On the other hand, most coaches expressed concerns about the capability of the chatbots, believing that it would be infeasible for them to replace face-to-face conversations fully:

It requires many skills to deal with highly human context [sic], not only listening to their words but also monitoring their behavior. Just asking template questions would never facilitate their reflection.

We understand this comment reflected the coach’s limited trust in the chatbot’s capability given the required skills in executive coaching as discussed in Section 2.1, although LLM-powered chatbots could go beyond simple template-based communication. Rather, they showed more positive attitudes to the chatbot assisting text communication between coaches and clients to reduce their burden and scale the coaching experience.

¹Here, we started the workshop by focusing on the coach side because we had difficulties inviting clients due to their contracts.

²<https://gpt4demo.com/apps/gpt4-playground>

I recognize the significance of text coaching in face-to-face sessions in maintaining relationships with clients and stimulating their actions. However, the difficulty lies in the fact that it could consume unlimited time, unlike face-to-face sessions with fixed schedules. AI coaches could adequately fulfill this role, possibly maintaining the quality and frequency of text communication.

Text communication between sessions is valuable to build a good foundation and to keep the momentum of their reflection, but in reality, it is not always possible for us to keep texting them, which could be counter-productive.

From these comments in the pre-use hearing regarding **RQ1**, we found that human coaches should keep playing a major role in facilitating clients' reflection to deal with diverse and highly interpersonal contexts in executive coaching. We also found that using chatbots as a text coach between face-to-face sessions to supplement them would be a promising direction to assist the coaches; although such communication is critical, it is a significant labor for the coaches, and clients could lose motivation by insufficient reply. Such a collaborative direction was further supported by the fact that coaches emphasized human coaches' importance for clients' behavior change motivation.

One of the reasons coaching is effective is the sense of being invested time by other people, so I believe fully automating with AI would be challenging.

Similar comments were obtained in the semi-structured interviews after the coaches interacted with a GPT-4-based chatbot.

The chatbot's response was better than I had expected. I believe it could be effective for people motivated to change their behavior independently, but for others, having a human coach to accompany them still seems crucial. For instance, if there is an opportunity to form a clear intention of how to and how often to use a chatbot coach beforehand through a dialogue with a human coach, then I think the client would achieve behavioral change.

These comments suggested that preparing clients' readiness for the chatbot coach is critical where they gain the motivation to keep using the system and the perspective regarding how it can foster their reflection. Moreover, the coaches agreed that human coaches should help the clients when they lose a way to utilize the chatbot coach.

I can envision cases where clients might stop using it midway due to the lack of a sense that someone is watching over them. The kind of relationship where a human coach provides comments or offers support when they are stuck will likely be necessary.

This comment emphasizes the importance of a human coach's presence for sustainable engagement, suggesting that it would not be optimal to fully automate the process even with common computational intervention approaches such as auto-reminders. The coaches concluded that their guidance on fostering the clients' readiness to

use chatbot coaches for behavioral change and maintaining functional relationships with chatbot coaches would be critical. This informed us of the answers to **RQ2**; blending coaching in collaboration of human and chatbot coaches can open up the way to providing successful coaching sessions in a scalable manner by reducing the workload of text coaching of human coaches.

5 EVALUATION STUDY: DESIGN AND METHOD

Based on the findings in the workshop, we prepared a chatbot-based prototype system for text coaching to empirically examine its influence on the clients in ecologically valid scenarios. We deployed a system in actual coaching scenes and tracked clients' reflections and interactions with the system over two weeks during July and August 2023. Finally, we conducted interviews with both coaches and clients to investigate the experience brought by the system qualitatively and to draw implications.

5.1 Prototype

Figure 1 presents the system we developed for the study. Informed by the workshop, this prototype focused on text coaching and was implemented to enable clients to have a text chat with the GPT-4-based³ chatbot coach (Figure 1A). The chat space is created for each text coaching session (Figure 1B). Clients can end the session (Figure 1C) and then automatically send a report to the human coach via email so that the coach can be involved in the clients' reflection process, as the workshop suggested. Here, the clients can revise the summary (Figure 1D) and messages (Figure 1E) to the human coach, where the summary is first generated by GPT-4. Moreover, we allow the clients to send a message to their human coaches in the middle of the text coaching with the chatbot coach (Figure 1F), which was guided by the comment in the workshop that the human coach would be required to help the client when needed. We also designed a prompt for the chatbot to suggest contacting the human coach if they struggle to maintain the conversation.

Importantly, we designed the prototype so that there would be a face-to-face session where a human coach would explain the use of the prototype and set a goal with the client before starting the chatbot-based text coaching. This is based on the important finding obtained in the workshop; that is, clients' readiness is critical to foster one's reflection using the system. We used this face-to-face session to set up their expectation during the trial period. More specifically, after explaining the capability of the chatbot coach, the human coach helped the client write their expected goal through the text coaching process and how many times they would use the system during the trial period. Note that the goal and usage expectation were set individually by each pair of the coach and client and were kept visible on the main page throughout the trial (Figure 1G).

The information was also used to create a prompt for the LLM inside the chatbot, as shown in Figure 2. Here, `{{ goal }}` and `{{ expectation }}` refer to the client's goal and expectation about the text coaching process, which were set during the face-to-face session and registered on the system after the session by the coach. This prompt was built on top of what we learned from the coaches'

³We used gpt-4-0613 with temperature = 0.

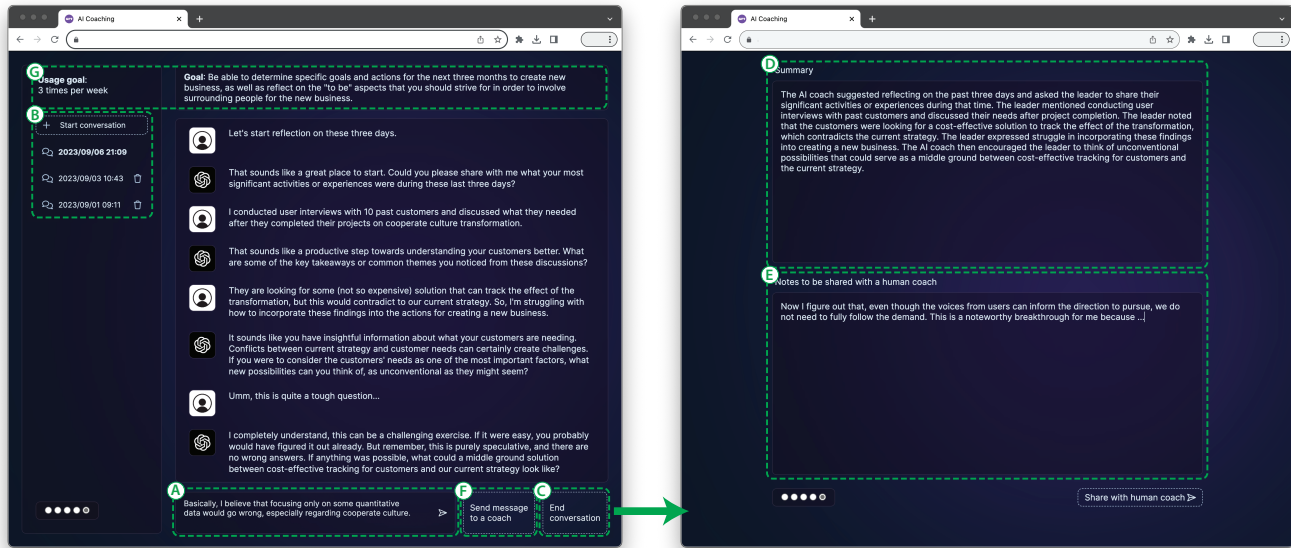


Figure 1: The prototype interface to use an LLM-powered chatbot coach. In the user study, this prototype was provided to clients and used at their own pace for two weeks to support their reflection to accomplish their professional goals. The presented conversation was derived from the use of a prototype by one of the authors to demonstrate its behavior.

You are a professional executive coach. Your role is to enhance the client's self-awareness and bring about behavioral change through precise questioning and feedback in your interactions with them.

The client's desired goals through coaching: {{ goal }}
What the client expects from you: {{ expectation }}

Please be mindful not to present multiple questions in a single interaction to avoid confusing the client. Furthermore, if you find it challenging to bring about behavioral change in your conversation with the client, please present the following message to them:
Speaking with a human coach might help clarify your thoughts further. What do you think about using the "Send message to a coach" button?

Figure 2: The prompt used in the chatbot guided by the workshop with coaches.

trials and errors in the workshop; for example, the sentences explaining the role were obtained from those the professional coaches had used to make the LLM imitate their conversations. As described above, we also allowed the LLM to suggest a client communicate with a human coach if needed. In addition, we observed that the LLM often produced multiple questions in a single response. According to the professional coaches, it was pointed out to lead to clients' confusion. Thus, we specified the LLM not to present multiple questions at once and finalized the prompt after several testing iterations.

5.2 Participants

We first recruited coaches from two coaching companies. We explained the concept of our study and asked them to introduce the prototype into their actual coaching with their clients. As a result,

ten pairs of coaches and their clients, who regularly have face-to-face sessions on a monthly or bi-weekly basis, participated in our study. We refer to clients as Cl1 – Cl10 and paired coaches as Co1 – Co10. The detailed background about the clients is summarized in Table 1.

5.3 Metrics

To elucidate how the system affects the client, we used questionnaire-based measurements and asked the clients to fill them out every time they used the prototype. In executive coaching, one's behavior or attitude to change shows a gradual transition, and we expected that investigating the transition using repeated questionnaires could provide insights. Given their labor, we focused on two aspects, i.e., the outcome of the reflection and attitude toward using the system, and used the corresponding two metrics.



Figure 3: The procedure of the user study using the developed chatbot text coach as a supplement to the regular coaching.

Table 1: Backgrounds of the clients who participated in the two-week user study.

	Gender	Age	History with the current coach
Cl1	M	50s	1 year
Cl2	F	40s	8 months
Cl3	F	40s	none
Cl4	F	40s	3 months
Cl5	F	30s	2 years
Cl6	M	40s	1.5 years
Cl7	F	30s	2 months
Cl8	F	50s	4 months
Cl9	F	40s	none
Cl10	F	20s	2 months

5.3.1 Authenticity Scale. There are a couple of approaches for assessing the quality of one’s reflection [11]. The authenticity scale [67] has been used in the context of executive coaching [5, 56] since it is about aligning one’s inner feelings and primary experiences with their external actions and communication, meaning that improving authenticity heavily relies on the depth of self-reflection. The scale consists of 12 items designed to measure the three factors: self-alienation, authentic living, and accepting external influence [7]. Self-alienation highlights the inherent discrepancy between an individual’s true experience and conscious awareness. A divergence between the two suggests the individual feels disconnected or unfamiliar with their true self. Authentic living assesses how well an individual’s actions and emotions align with their internal state awareness. Accepting external influence reflects an individual’s propensity to be swayed by others, representing the impact of their social surroundings. In the questionnaires for this scale, participants respond to each item using a seven-point Likert scale, ranging from 1 (“does not describe me at all”) to 5 (“describes me very well”).

5.3.2 Behavioral Intention. Behavioral intention was prepared to evaluate the clients’ attitudes toward using the system. The Technology Acceptance Model [17] guides the concept of behavioral intention, which explains users’ attitudes towards technologies and is frequently used to evaluate how likely individuals are to use the technologies. As we confirmed in the workshop (Section 4.2), clients’ readiness is key to the outcome of their experience of text

coaching. Moreover, their intention to use the system depends on the quality of the experience. Therefore, we thought this metric would provide useful context for analyzing the qualitative results of the semi-structured interviews. We used the questionnaire from a previous study [59] to measure behavioral intention. This questionnaire consisted of three questions assessing respondents’ intentions to continue using a system, with responses scored on a scale from 1 to 5.

5.4 Procedure

Figure 3 shows the procedure of our study. First, an experimenter explained the prototype to the coaches by showing examples and having them play the system for a while. Then, the coach conducted a face-to-face session with the client to explain the system’s use and set the text coaching goal, which took roughly 30 minutes. Specifically, they discussed their current issue, how they would utilize the chatbot, the goal state after the two weeks, and how often they planned to interact with the chatbot; this phase was informed by the workshop. At the end of the session, the clients sent the answers to these questions on Google Form. The answers were then used for configuring the system, as described in Section 5.1. This form also included questionnaires for behavioral intention. Then, in the next two weeks, the client used the text-coaching system at their own pace; we did not set a dedicated time, nor did we remind them to do it, to preserve ecological validity. After each text coaching, they were navigated to answer questionnaires about the authenticity scale and behavioral intention as well as share any comments about the experience. During the two-week trial period, there were no face-to-face sessions between the coach and the client. After this period, the client and coach reflected on their text-coaching experience, discussing how they used the system and whether it met their goals. Lastly, after the session, the experimenter conducted semi-structured interviews with the coach and the client individually, which took roughly 30 minutes. For the coaches, they asked a series of questions: “Compared to your usual coaching, how would you describe the coaching experience made possible by the AI technology?”, “What are the advantages and disadvantages of chatbot coach?”, and “Would you like to keep using the system in your coaching, and if yes, how?”. On the other hand, the clients were asked about the following: “How did your behavior change during the coaching period?”, “Were you able to achieve the goal you had set in the beginning?”, “How would you change your use of the system if you did it again?”, “How has your trust in the chatbot

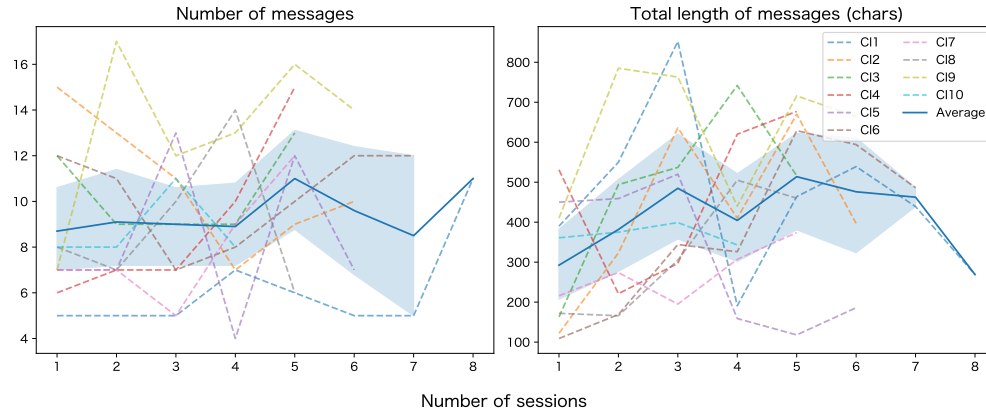


Figure 4: (Left) The number of messages each client sent to the chatbot coach per session. (Right) The total length of the messages per session in the number of characters. The blue area highlights indicate the 95% confidence interval of the average value.

coach changed over time?”, and “Would you like to keep using the system, and if yes, how?”.

6 RESULTS

6.1 Chat Behavior

We first examined how the clients used the chatbot coach (See Figure 4). During the two-week trial period, the average number of text-coaching sessions was 5.7 ($SD = 1.1$). The variance is understandable, given the expected number had been arbitrarily set during the first face-to-face session. Note that, although we did not remind them, all participants except one client met the declared number. Additionally, while there was a variance in the chat behavior by participants and sessions, their average number of messages and their lengths were sustained over the sessions. The results suggested that the participants engaged in text coaching throughout the study.

Figure 5 provides an actual chat history of one client, which was presented with their permission. This shows that the client promoted her reflection within 10–15 messages. Also, we can infer that the responses from the chatbot coach motivated the client to take action towards her goal, as the chatbot effectively utilized the coaching skill of “acknowledgment” [63]. Specifically, coaches are known to induce the client’s new behavior by acknowledging a client’s action and, more importantly, reflecting on how that action honored their values. To understand their experience of using the chatbot coach and inform the answer for **RQ3**, we analyzed their comments during our semi-structured interviews in the next section.

6.2 Semi-Structured Interviews

The participants’ (i.e., both coaches’ and clients’) responses in the semi-structured interviews were analyzed using open coding [55]. Through iterative refinement processes, we obtained four topics, as follows.

6.2.1 Power of a Chatbot Coach as a Promoter of Clients’ Actions When Human Coach Is Not Present. Firstly, we found that all clients responded affirmatively about the power of the chatbot in promoting their actions, like:

Being able to communicate with the coach at my preferred time was casual and nice. This was helpful in maintaining my motivation. (CI18)

Similar comments were also obtained from the coach side, such as:

It was quite convenient that the coaching proceeded without the trouble of fixing the session schedule in advance. (Co6)

This confirms the discussions we observed in the workshop. Moreover, the clients’ comments shed light on how their reflection was prompted through the dialogue,

By being repeatedly and persistently asked, like “then, what exactly will the next step you do?”, I was able to clarify what kind of preparation and planning was necessary. (CI2)

This advantage that the chatbot can help clients break down actions underpins our prompt design guided by the workshop. We believe this has been made possible by the LLM’s zero-shot reasoning capability in diverse problem-solving tasks [35].

In addition, some clients sympathized with the chatbot more than we had expected.

Even receiving simple replies like “that’s good” made me happy, even though I knew the opponent was an AI. I found it surprisingly delightful, and it helped me make progress. (CI1)

This point helped the clients take practical action based on their reflection, confirming the effect of the chatbot’s acknowledgment that we discussed in Section 6.1. These comments suggest the power of LLM-powered chatbots in inducing clients’ reflection and sustainable engagement for behavioral changes.

Client's goal:	Being able to constantly think ahead three days and take action, thereby creating both temporal and mental flexibility, allowing me to focus on emotions, both within myself and with others
Client:	I'm going to have a reflection on yesterday.
Chatbot:	Understood. Could you please tell me about your yesterday specifically if you were able to think about things three days ahead and take actions towards that?
Client:	Until yesterday, I had been taking a four-day course on management and leadership at graduate school. During the day, I mostly dedicated my time to the course, so I did the preparation for work three days ahead during my free time, early in the morning, and before bedtime.
Chatbot:	Taking action to prepare for work three days ahead while attending graduate school is a commendable effort and strategy in itself. Among these specific actions, what do you feel was particularly effective?
Client:	Keeping a to-do list to make sure I don't forget anything. Also, trying to communicate with others as timely as possible. I also tackled tasks requiring creativity in the early morning.
Chatbot:	I see, it seems like you have a good understanding of effective approaches for each activity. Knowing your peak productivity times and priorities and effectively managing your tasks is commendable. What specific goals do you want to achieve in the next three days?
Client:	Since the Bon (summer) holiday is coming up, I want to finish gathering information on matters that require confirmation with others before the holiday starts.
Chatbot:	That's a good goal. To relax during the Bon holiday, it's essential to complete tasks that require communication with others in advance. What are your specific plans to achieve this goal?
Client:	I will complete tasks that require communication as much as possible until the day before the Bon holiday.
Chatbot:	That's a strong commitment. Organizing everything and entering the holiday without unnecessary burdens will have a positive impact on your mental well-being. To achieve this, which tasks do you think you should start with?
Client:	It's the preparations for upcoming negotiations and presentations right after the Bon holiday.
Chatbot:	That's a clear plan. By prioritizing important tasks like negotiations and presentations, you'll likely have a more relaxed Bon holiday. You probably already have specific action plans for each task in mind. Do you have those plans ready?
Client:	Yes.
Chatbot:	Great, that shows your plans are well-structured. Having a clear vision of what to do in the coming days brings you closer to achieving your goals. If there are any potential challenges when executing your plans, think about how to overcome them.
Client:	There are no specific challenges, just that I want to review what I've learned during these four days at graduate school while managing these tasks. It might add to the workload, but I'll definitely get it done.
Chatbot:	Reviewing what you've learned at graduate school is important. It might make your schedule busier, but with your determination and planning, I believe you'll successfully complete your tasks. What are your specific plans for the review process?
Client:	I plan to reflect, articulate, and organize my learnings.
Chatbot:	That approach is highly effective. By going through the steps of reflection, articulation, and organization, you can deepen your understanding and absorb your new knowledge and skills. Do you already have a specific action plan for this?
Client:	Yes, I will allocate time for this in the early morning. I will create outputs of my learnings.
Chatbot:	That plan will help solidify your newly acquired knowledge. Allocating time in the early morning to create outputs of your learnings will allow you to apply what you've learned throughout the day. Having a clear action plan like this is excellent. Please proceed with your plan for tomorrow, and regardless of the outcome, it will undoubtedly contribute to your personal growth. Have a great day tomorrow!
Client:	Due to time constraints, I'll end here for today, is that okay?
Chatbot:	Of course. Thank you for taking the time to discuss your plans today. Sharing your thoughts and action plans shows your commitment to your growth and goal achievement. I look forward to hearing about your progress next time. Have a wonderful day!
Client:	Thank you. Have a great day!
Chatbot:	You're welcome. Your proactive attitude was impressive. Have a good day, and I look forward to your next contact!

Figure 5: Example from one client's messages with the chatbot coach. It is observed that the chatbot often acknowledges the client's actions and asks questions that can further break down the problems they face.

6.2.2 Limitation of a Chatbot Coach in Inducing Deep Reflection. At the same time, the specific aspect of the chatbot that can be derived from the nature of LLMs appeared to be a limitation.

As the conversation continued, I found it would be beneficial if the chatbot asked in-depth questions like "Isn't the goal you set initially a bit lenient?" or "What's the true significance of pursuing this goal?" (C17)

Questions like identifying behaviors the client unconsciously avoids would deepen the conversation but were not observed. I believe that such aspects would not been covered unless I proactively intervened in the conversation as a human's role. (Co2)

This would be intrinsic to the LLM we used, GPT-4, which is trained to follow users' intention [48] so that it would not generate exceptional responses. On the other hand, to induce deep reflection of clients, coaching sometimes requires questions that challenge the clients and would make them uncomfortable [63]. This point can be a limitation of chatbot coaches, conversely suggesting the necessity of human coaches' involvement in the process.

We also would like to note that the button to ask human coaches for guidance during the text coaching sessions (Figure 1F) was rarely used. In this regard, one client commented:

I did not have much opportunity to use the button because I was satisfied with the fact that some actions were progressing even without deep, challenging questions. (C18)

One coach also mentioned:

Estimating the timing to interject a sharp retort or comment is very difficult even for a human coach, and it's very important. Personally, I think there is a need to guide the conversation from our end, rather than through text, but through face-to-face conversation. (Co10)

This would imply the need for human coaches to carefully monitor the communication between clients and the chatbot coach, as initiating such communication from the client side would be challenging.

6.2.3 Importance of Clarifying Goals and Securing Commitment with a Human Coach. Despite such limitations, the clients favored the overall experience of the text coaching, maintaining the engagement during the trial as we observed in Section 6.1. Their comments revealed that this was due to the designed approach of blending human and chatbot coaches, even though they were not informed about the discussion results of the previous workshop.

Making an initial commitment with the human coach regarding the extent of using the chatbot became the motivation for actually following through with it. (Cl4)

The fact that the human coach was keeping track of my progress positively motivated me to engage in text coaching with the chatbot. (Cl3)

A similar perspective was provided also by the coaches.

Once we set a goal and made a promise about the usage with solid motivation, the entire process proceeded without any effort. This would be one of the optimal ways to introduce the chatbot to executive coaching. (Co6)

These comments confirm the effectiveness of our design in promoting the blended initiatives of the human and chatbot coaches, which was informed by the workshop.

Moreover, we found that the blended initiative can foster the self-disclosure of the clients.

I can flatly talk to the chatbot. With a human coach, it takes time to realize trust, and sometimes it's a little difficult to be honest. The chatbot was an easy contact of communication between the human coach and me. (Cl9)

Some people need a long time to disclose themselves to coaches. But, since the chatbot can provide frequent communication, their disclosure can be accelerated. I learned that the frequency of the text coaching sessions can be a key to successful outcomes. (Co4)

These comments were particularly interesting because it was the opposite of the expectation of the literature [57] that suggested the difficulty of clients in trusting AIs. We suppose this would not be realized without the human coaches who secure the commitment of the clients' frequent chatbot use.

6.2.4 Direction for the Improved Blending Coaching. Given such observations and limitations, some coaches suggested directions for enabling better collaboration between clients and chatbot coaches.

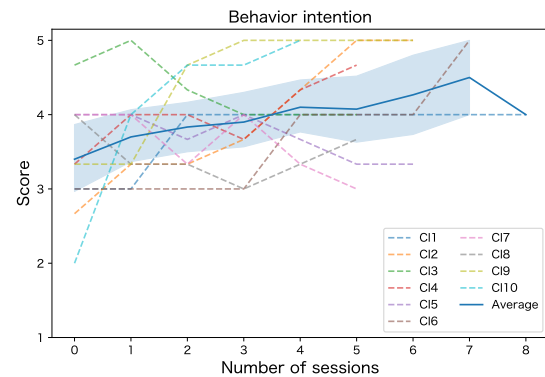


Figure 6: The transition of the score of the participated clients' behavioral intention to use the chatbot coach. The blue area highlights the 95% confidence interval of the average score.

It would be good to allocate 60 minutes for the first session before starting to use the chatbot coach to confirm the significance of the goal a client sets and to adjust the difficulty level of the goal. This would reduce the risk of clients working on goals that are not essential or too easy. (Co4)

When the client did not seem to be making good progress, it would have been better to have an opportunity to talk for five minutes or so instead of communicating via text. When I was writing my reply to the client in text, I thought this text could also be generated by an AI with some tuning. Conversely, I felt that the value of a human coach lies in the ability to ask tough questions in person, taking into consideration the subtleties in clients' speech and other aspects. (Co8)

We believe that the values of our study are not limited to confirming the effectiveness of our design but include these practical insights provided by coaches who actually experienced the blended approach. Later in Section 7, we summarize our findings to foster the effective use of LLM-powered chatbots in executive coaching.

6.3 Clients' Authenticity Scale and Behavioral Intention

We also analyzed the clients' evaluations of their authenticity and behavioral intention, which were collected after the end of each text coaching session. Figure 6 shows the transition of their scores of the behavioral intention, in which we did not find negative responses once they started using the chatbot coach. Rather, we found many participants who increased their scores as they had more sessions, except for two clients. We inferred that their consistent use of the chatbot, as discussed in Section 6.1, resulted from their positive acceptance. At the same time, we confirmed that the two clients who showed a decreasing trend mentioned the limitations of the chatbot during the semi-structured interviews, as described in Section 6.2.2. This conversely implied the importance of human

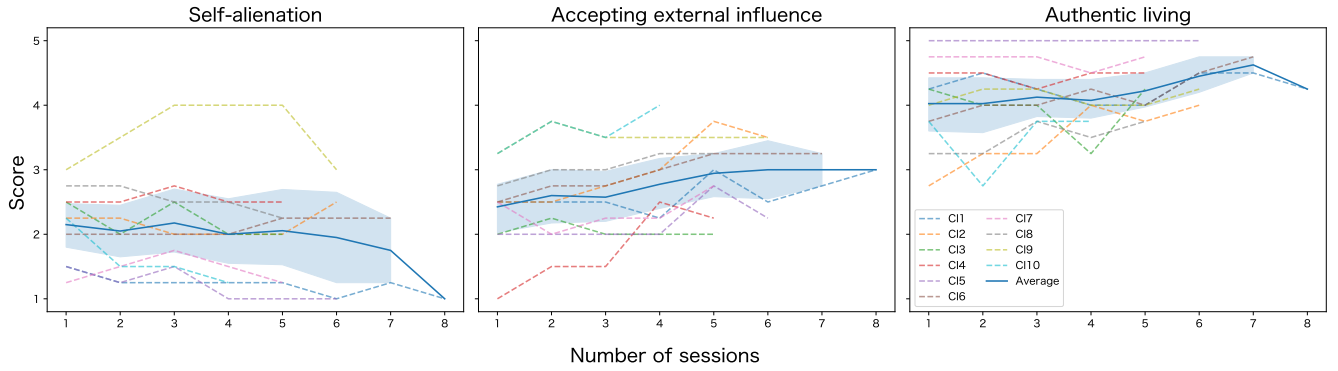


Figure 7: The transitions of the scores of the three factors of the participated clients' authenticity. The blue area highlights the 95% confidence interval of the average score.

coaches' involvement; clients' behavioral intentions might decay without such involvement, especially when they need deeper reflection.

Figure 7 shows the transitions of the clients' authenticity scores. We observed the most dynamic transition regarding accepting external influence, reflecting that the clients started to change their behavior based on the conversations with the chatbot coach. Given that such actions functioned toward achieving their declared goals (see Section 6.2.1), this suggests the effectiveness of the chatbot coach in inducing positive outcomes through text coaching. Also, we can see a decreasing trend in self-alienation. Though it was not so significant, we conjectured that reflection through dialogue with the chatbot coach could enhance their conscious awareness. At the same time, given the observed limitations of the chatbot (see Section 6.2.2), the effect could be further enhanced by exploring the optimal blending of human and chatbot coaches, for instance, as suggested by Co8 as an improvement direction.

In conclusion, our findings for **RQ3** revealed that the chatbot coach significantly contributed to fostering clients' reflection and inspiring their actions, especially in the absence of a human coach. Furthermore, the integrated design of human and chatbot coaches, as derived from insights in **RQ2**, proved effective, encouraging consistent client engagement with the chatbot. The limitations identified, inherent to current LLMs, underscored the necessity of this blended approach, which enables clients to have deep self-reflection effectively. The outcomes associated with **RQ3** highlight the potential to broaden the reach of executive coaching, making it accessible to all those in need through the support of LLMs.

7 DISCUSSION

Through building a prototype with insights from coaches and conducting an empirical study in actual coaching scenes, we have shed light on the plausible form of using LLM-powered chatbots effectively for one's leadership growth. Lastly, we discuss the study's implications for the HCI community and derive a guideline for deploying such chatbots in HRD practice.

7.1 Types of Reflections Chatbot Coaches Support

The results of the semi-structured interviews suggested that the LLM-powered chatbot coach facilitated clients' reflection while uncovering their limitations. This ties into the discussion on the contrasting learning concepts, single-loop learning and double-loop learning [6]. Single-loop learning involves adjusting and correcting existing frameworks and policies to address discrepancies without changing underlying assumptions or values. In contrast, double-loop learning goes further by analyzing and potentially altering the fundamental assumptions or governing values, facilitating more profound organizational changes and improvements. This scheme clarifies the advantages and disadvantages of introducing the LLM-powered chatbot coach. First, our results suggest that the chatbot coach can support clients' single-loop learning by iteratively asking questions to clarify the steps toward the set goal while acknowledging their actions, as shown in Figure 5. In our study, the human coach and the client favored the benefit of chatbots always being available. This results not only in reducing the workload of human coaches but also in fostering the reflection of clients on their convenient occasions. However, as discussed in Section 6.2.2, it is challenging to facilitate their double-loop learning because posing incisive and pushing questions can be regarded as too sensitive for the chatbot coach. Specifically, we need to acknowledge that making an LLM-powered chatbot capable of asking challenging questions increases the risk of manipulating clients' emotions in a harmful manner, as Cabrera *et al.* [13] discussed in the context of why chatbots need human supervision. Therefore, while Liu *et al.* [39] reported the effectiveness of introducing the idea of double-loop learning into skill training, it is also suggested that automating such learning processes is sometimes difficult, especially when deep reflection is required.

7.2 Roles of Human Coach in Executive Coaching Dialogue

Our study also revealed the advantages and disadvantages of human coaches, especially through the lens of single/double-loop learning. For example, we can infer from the obtained comments that the

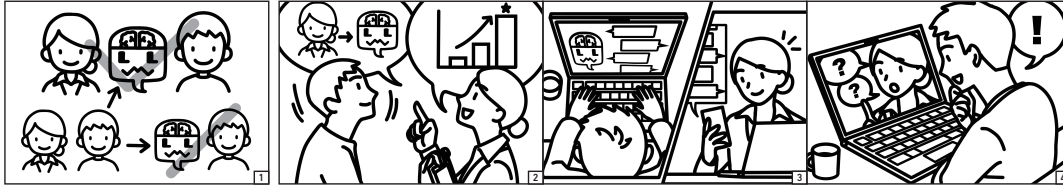


Figure 8: Guideline for the blended approach of human and LLM-powered chatbot coach for leadership growth in executive coaching, synthesized through our study. The human coach is suggested to follow the steps to introduce the chatbot coach successfully and to augment clients' reflection: 1) introducing the chatbot as complementary text coaching, 2) conducting pre-session to foster a client's readiness before initiating chatbot-driven coaching, 3) maintaining a communication channel with the client during the text coaching period, and 4) monitoring client progress and identify moments to intervene.

human coaches were crucial for inducing clients' deep reflection through potentially uncomfortable inquiry. Prior work suggested that professional coaches can interpret subtle signals of clients, such as nonverbal behavioral cues during face-to-face conversation, to pose such questions [4]. Our study also suggested that the existence of human coaches can contribute to clients maintaining their engagement with the entire coaching experience. However, human coaches are not always available, which limits the opportunity for clients to have reflection instantly; as discussed in Section 2, the industry of executive coaching faces the problem of lacking proficient coaches. Together with the advantages and disadvantages of chatbot coaches, the effective blending of both parties is demanded, in contrast to focusing on the engagement with one side, and our study also uncovered multiple factors key to achieving it, which we summarize in the next section.

7.3 Toward Deployment of Chatbot Coach for Leadership Growth

To conclude this paper, we summarize the guided flow for human coaches to follow to blend human and chatbot coaches in executive coaching effectively, the overview of which is presented in Figure 8.

7.3.1 Introduce Chatbot as Complementary Text Coaching. The workshop indicated that delegating all components human coaches take in executive coaching to chatbots is not reasonable, given the significance of actual human presence in influencing clients' behavior. On the other hand, our study confirmed chatbots' effectiveness in reducing the workload of text coaching to keep the clients' attitude to improve their behavior. Thus, clarifying the point that chatbots serve as a complement to traditional face-to-face sessions is important.

7.3.2 Conduct Pre-Session Before Introducing Chatbot Coach. Our study also suggested the importance of setting a clear goal for text coaching to ensure a successful experience. In detail, due to the nature of chatbots, efficient usage naturally forms within the client through use when their readiness is enough. Thus, the human coach is encouraged to spend a certain amount of time with the client to adjust the difficulty level of the goal and expected outcome, which would last 30–60 minutes.

7.3.3 Maintain Communication Channel after Text Coaching. Furthermore, the presence of the human coach should be conveyed

during the text coaching period to keep the client's motivation and readiness. A communication channel after text coaching, e.g., the client sending its summary to the human coach and the coach leaving reactions to it, would be recommended.

7.3.4 Monitor Client Progress and Identify Moments to Intervene Critically. Our study confirmed that the intervention from the human coach is desired to induce deeper reflection, e.g., double-loop learning. Currently, the best way to have such opportunities remains an open question. As implied in Section 6.2.2, it would be hard for clients themselves to initiate such conversation as it is by nature mentally challenging. Therefore, the coach needs to monitor their progress by, for instance, paying attention to the summary report the clients send or introducing a quick survey after the text coaching to ask about their satisfaction.

7.4 Implication for Chatbot-Based Reflection Support in Other Settings

The advantage of the chatbot coach we discussed in Section 6.2.1 shares aspects with that of the chatbots in other domains as discussed in Section 2.3, that is, being available when the human coach is not present and providing consistent feedback or answers. Furthermore, the recent advancement in LLM enables chatbots to ask natural and reasonable questions to foster actions to achieve the planned goal, showing a greater capability than a previously-explored chatbot, e.g., scripted feedback of the health coach in [43]. Meanwhile, in the context of executive coaching, the importance of nurturing a space for critical introspection has been highlighted, and we conclude that a blend of human and chatbot coaches is necessary. Such a design can apply to reflection support in other domains, where users need to be rigorously challenged by external sources to transition into a more ideal state. For example, Li *et al.* [38] explored the design of LLM-powered chatbots to encourage moderate smartphone use and emphasized the importance of personalized dialogue. This work can be augmented by our exploration into a complementary strategy, namely, the blended form of human and chatbot coaches. In addition, Lee *et al.* [37] showed that, for mental health care, not only providing care from a chatbot but also asking users to provide care for a (vulnerably conditioned) chatbot can be effective. This practice can be used in a blended manner, in which human therapists contribute to clarifying the goal of the

interaction with a chatbot and maintaining the motivation to continue caring for the chatbot. Our empirical findings highlight an emerging theme that demands this blended approach, especially in the era of rapidly evolving LLMs.

8 CONCLUSION AND FUTURE WORK

We clarified the potential of an LLM-powered chatbot to foster one's deep reflection to achieve leadership growth by taking executive coaching as a study field. Our studies in actual coaching scenes revealed the strength of both human coach and chatbot coach and synthesized a way to combine them to realize optimal reflection experience. The current study involves a limited number of coaching organizations, which might result in homogeneous samples sharing a single cultural context. Thus, more participants (especially coaches) at different organizations are expected to verify the findings further.

In addition, since individual coaching experience varies significantly, conducting a comparison study with a baseline condition was difficult without involving a large participant pool. Rather, we co-designed with both coaches and clients to advance the practice with an emerging technology. Again, a larger user study with a baseline condition (e.g., dropping one of the key steps we identified in Section 7.3) will bring additional evidence to quantify the effect. It would be beneficial if we could conduct the study in the long term because it would reduce the possibility that a novelty effect occurred, as pointed out by Weber *et al.* [61].

Discussion on ethical considerations is also crucial to push the deployment of the chatbot forward. For example, clarifying the role of the chatbot and the scope of information sharing when introducing the chatbot (see Section 7.3.1) would be crucial in terms of both data privacy and psychological safety. Here, the standard in executive coaching about confidentiality [27] can also be applied to this blended approach. At the same time, the necessity of monitoring the risk of introducing bias or discrimination in chatbots [20] is common to us, while we believe that our design of trying not to personify the chatbot can be a remedy.

Furthermore, we can extend the blended approach to develop a socio-technical tool that allows practitioners to leverage the approach in other domains, as Sadek *et al.* [53] suggested. At the same time, we need to be aware that the outcome of chatbot coaches may evolve with the future advancement of LLMs. Nevertheless, we believe that our qualitative data show the enduring value of human coaches and consistently support the benefits of the blended approach.

ACKNOWLEDGMENTS

This work was supported in part by JST ACT-X Grant Number JPMJAX200R and JSPS KAKENHI Grant Numbers JP21J20353. Also, this work was conducted in collaboration with and under the supervision of Teambox, Inc.

REFERENCES

- [1] Deemah A. Alqahtani, Caroline Jay, and Markel Vigo. 2020. The Effect of Goal Moderation on the Achievement and Satisfaction of Physical Activity Goals. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4 (2020), 116:1–116:18. <https://doi.org/10.1145/3432209>
- [2] Saleema Amershi, Daniel S. Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi T. Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 ACM CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 3. <https://doi.org/10.1145/3290605.3300233>
- [3] Riku Arakawa, Kiyosu Maeda, and Hiromu Yakura. 2024. Supporting Experts with a Multimodal Machine-Learning-Based Tool for Human Behavior Analysis of Conversational Videos. *CoRR* abs/2402.11145 (2024). <https://doi.org/10.48550/ARXIV.2402.11145>
- [4] Riku Arakawa and Hiromu Yakura. 2019. REscUE: A framework for REal-time feedback on behavioral CUEs using multimodal anomaly detection. In *Proceedings of the 2019 ACM CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 572. <https://doi.org/10.1145/3290605.3300802>
- [5] Riku Arakawa and Hiromu Yakura. 2020. INWARD: A computer-supported tool for video-reflection improves efficiency and effectiveness in executive coaching. In *Proceedings of the 2020 ACM CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–13. <https://doi.org/10.1145/3313831.3376703>
- [6] Chris Argyris. 1977. Double loop learning in organizations. *Harvard business review* 55, 5 (1977), 115–125.
- [7] Godfrey T Barrett-Lennard. 1999. *Carl Rogers' helping system: Journey and substance*. SAGE Publications, London, UK.
- [8] Eric P. S. Baumer, Vera D. Khovanskaya, Mark Matthews, Lindsay Reynolds, Victoria Schwanda Sosik, and Geri Gay. 2014. Reviewing reflection: On the use of reflection in interactive system design. In *Proceedings of the 2014 ACM Designing Interactive Systems Conference*. ACM, New York, NY, 93–102. <https://doi.org/10.1145/2598510.2598598>
- [9] Jeremy Beaudry, Alyssa Consigli, Colleen Clark, and Keith J. Robinson. 2019. Getting ready for adult healthcare: Designing a chatbot to coach adolescents with special health needs through the transitions of care. *Journal of Pediatric Nursing* 49 (2019), 85–91. <https://doi.org/10.1016/j.pedn.2019.09.004>
- [10] Ivo Benke, Sebastian Vetter, and Alexander Maedche. 2021. LeadBoSki: A Smart Personal Assistant for Leadership Support in Video-Meetings. In *Companion Publication of the 2021 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 19–22. <https://doi.org/10.1145/3462204.3481764>
- [11] Marit Bentvelzen, Pawel W. Wozniak, Pia S. F. Herbes, Evropi Stefanidi, and Jasmin Niess. 2022. Revisiting reflection in HCI: Four design resources for technologies that support reflection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 2:1–2:27. <https://doi.org/10.1145/3517233>
- [12] James Bridgeman and Andrea Giraldez-Hayes. 2023. Using artificial intelligence-enhanced video feedback for reflective practice in coach development: Benefits and potential drawbacks. *Coaching: An International Journal of Theory, Research and Practice* (2023), 1–18. <https://doi.org/10.1080/17521882.2023.2228416>
- [13] Johana Cabrera, Maria Soledad Loyola, Irene Magaña, and Rodrigo Rojas. 2023. Ethical Dilemmas, Mental Health, Artificial Intelligence, and LLM-Based Chatbots. In *Proceedings of the 10th International Work-Conference on Bioinformatics and Biomedical Engineering*, Vol. 13920. Springer, 313–326. https://doi.org/10.1007/978-3-031-34960-7_22
- [14] Wanling Cai, Yucheng Jin, Xianglin Zhao, and Li Chen. 2023. "Listen to music, listen to yourself": Design of a conversational agent to support self-awareness while listening to music. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 119:1–119:19. <https://doi.org/10.1145/3544548.3581427>
- [15] Jacky Casas, Elena Mugellini, and Omar Abou Khaled. 2018. Food diary coaching chatbot. In *Proceedings of the 2018 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Wearable Computers*. ACM, New York, NY, 1676–1680. <https://doi.org/10.1145/3267305.3274191>
- [16] Mara Castro Correia, Nuno Rebelo dos Santos, and Jonathan Passmore. 2016. Understanding the Coach-Coachee-Client relationship: A conceptual framework for executive coaching. *International Coaching Psychology Review* 11, 1 (2016), 6–23.
- [17] Fred D. Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly* 13, 3 (1989), 319–340. <https://doi.org/10.2307/249008>
- [18] Yogesh K. Dwivedi et al. 2023. Opinion paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management* 71 (2023), 102642. <https://doi.org/10.1016/j.ijinfomgt.2023.102642>
- [19] Harry Barton Essel, Dimitrios Vlachopoulos, Akosua Tachie-Menson, Esi Eduafua Johnson, and Papa Kwame Baah. 2022. The impact of a virtual teaching assistant (chatbot) on students' learning in Ghanaian higher education. *International Journal of Educational Technology in Higher Education* 19, 1 (2022). <https://doi.org/10.1186/s41239-022-00362-6>
- [20] Jasper Feine, Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2019. Gender Bias in Chatbot Design. In *Proceedings of the 3rd International Workshop on Chatbot Research and Design*. Springer, 79–93. https://doi.org/10.1007/978-3-030-39540-7_6

- [21] Daniel C. Feldman and Melenie J. Lankau. 2005. Executive coaching: A review and agenda for future research. *Journal of Management* 31, 6 (2005), 829–848. <https://doi.org/10.1177/0149206305279599>
- [22] Asbjørn Følstad, Marita Skjue, and Petter Bae Brandtzaeg. 2019. Different chatbots for different purposes: Towards a typology of chatbots to understand interaction design. In *Proceedings of the 2019 International Conference on Internet Science*. Springer International Publishing, Cham, Switzerland, 145–156. https://doi.org/10.1007/978-3-030-17705-8_13
- [23] Corey Ford and Nick Bryan-Kinns. 2023. Towards a reflection in creative experience questionnaire. In *Proceedings of the 2023 ACM CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 763:1–763:16. <https://doi.org/10.1145/3544548.3581077>
- [24] Silvia Gabrielli, Silvia Rizzi, Sara Carbone, and Valeria Donisi. 2020. A chatbot-based coaching intervention for adolescents to promote life skills: Pilot study. *JMR Human Factors* 7, 1 (2020), e16762. <https://doi.org/10.2196/16762>
- [25] Anthony M. Grant. 2012. Making positive change: A randomized study comparing solution-focused vs. problem-focused coaching questions. *Journal of Systemic Therapies* 31, 2 (2012), 21–35. <https://doi.org/10.1080/105218807.2012.31.2.21>
- [26] Carolin Graßmann and Carsten C Schermuly. 2021. Coaching with artificial intelligence: Concepts and capabilities. *Human Resource Development Review* 20, 1 (2021), 106–126.
- [27] Daniel P. Greenfield and William K. Hengen. 2004. Confidentiality in Coaching. *Consulting to Management* 15, 1, 9–14.
- [28] Peter Hawkins. 2008. The coaching profession: some of the key challenges. *Coaching: An International Journal of Theory, Research and Practice* 1, 1 (2008), 28–38. <https://doi.org/10.1080/175218807.01878174>
- [29] Baek-Kyoo (Brian) Joo. 2005. Executive Coaching: A Conceptual Framework From an Integrative Review of Practice and Research. *Human Resource Development Review* 4, 4 (2005), 462–488. <https://doi.org/10.1177/1534484305280866>
- [30] Matthew Jørke, Yasaman S. Sefidgar, Talie Massachi, Jina Suh, and Gonzalo A. Ramos. 2023. Pearl: A technology probe for machine-assisted reflection on personal data. In *Proceedings of the 28th ACM International Conference on Intelligent User Interfaces*. ACM, New York, NY, 902–918. <https://doi.org/10.1145/3581641.3584054>
- [31] Mira El Kamali, Leonardo Angelini, Maurizio Caon, Giuseppe Andreoni, Omar Abou Khaled, and Elena Mugellini. 2018. Towards the NESTORE eCoach: A tangible and embodied conversational agent for older adults. In *Proceedings of the 2018 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Wearable Computers*. ACM, New York, NY, 1656–1663. <https://doi.org/10.1145/3267305.3274188>
- [32] Richard R. Kilburg. 1997. Coaching and executive character: Core problems and basic approaches. *Consulting Psychology Journal: Practice and Research* 49, 4 (1997), 281–299. <https://doi.org/10.1037/1061-4087.49.4.281>
- [33] Wonjung Kim, Seungchul Lee, Seonghoo Kim, Sungbin Jo, Chungkuk Yoo, Inseok Hwang, Seungwoo Kang, and Junehwa Song. 2020. Dyadic Mirror: Everyday Second-person Live-view for Empathetic Reflection upon Parent-child Interaction. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3 (2020), 86:1–86:29. <https://doi.org/10.1145/3411815>
- [34] Rafal Kocielnik, Lillian Xiao, Daniel Avrahami, and Gary Hsieh. 2018. Reflection Companion: A Conversational System for Engaging Users in Reflection on Physical Activity. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 2 (2018), 70:1–70:26. <https://doi.org/10.1145/3214273>
- [35] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. In *Proceedings of the 36th Annual Conference on Neural Information Processing Systems*. Curran Associates, Inc., Red Hook, NY, 22199–22213.
- [36] Mohammad Amin Kuhail, Nazik Alturki, Salwa Alramlawi, and Kholood Alhejori. 2023. Interacting with educational chatbots: A systematic review. *Education and Information Technologies* 28, 1 (2023), 973–1018. <https://doi.org/10.1007/s10639-022-11177-3>
- [37] Minha Lee, Sander Ackermans, Nena van As, Hanwen Chang, Enzo Lucas, and Wijnand A. IJsselstein. 2019. Caring for Vincent: A Chatbot for Self-Compassion. In *Proceedings of the 2019 ACM CHI Conference on Human Factors in Computing Systems*. ACM, 702:1–702:13. <https://doi.org/10.1145/3290605.3300932>
- [38] Zhuoyang Li, Minhui Liang, Hai Trung Le, Ray LC, and Yuhua Luo. 2023. Exploring Design Opportunities for Reflective Conversational Agents to Reduce Compulsive Smartphone Use. In *Proceedings of the 5th International Conference on Conversational User Interfaces*. ACM, 37:1–37:6. <https://doi.org/10.1145/3571884.3604305>
- [39] Chenchen Liu, Jierui Hou, Yun-Fang Tu, Youmei Wang, and Gwo-Jen Hwang. 2021. Incorporating a Reflective Thinking Promoting Mechanism into Artificial Intelligence-Supported English Writing Environments. *Interactive Learning Environments* 31, 9 (2021), 5614–5632. <https://doi.org/10.1080/10494820.2021.2012812>
- [40] Doug MacKie. 2015. The effects of coachee readiness and core self-evaluations on leadership coaching outcomes: A controlled trial. *Coaching: An International Journal of Theory, Research and Practice* 8, 2 (2015), 120–136. <https://doi.org/10.1080/17521882.2015.1019532>
- [41] Vanessa Mai, Annika Wolff, Anja Richert, and Ivonne Preusser. 2021. Accompanying reflection processes by an AI-based StudiCoachBot: A study on rapport building in human-machine coaching using self disclosure. In *Proceedings of the 23rd HCI International Conference - Late Breaking Papers*. Springer, Cham, Switzerland, 439–457. https://doi.org/10.1007/978-3-030-90328-2_29
- [42] Elliot G. Mitchell, Noemie Elhadad, and Lena Mamykina. 2022. Examining AI methods for micro-coaching dialogs. In *Proceedings of the 2022 ACM CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 440:1–440:24. <https://doi.org/10.1145/3491102.3501886>
- [43] Elliot G. Mitchell, Rosa Maimone, Andrea Cassells, Jonathan N. Tobin, Patricia G. Davidson, Arlene M. Smaldone, and Lena Mamykina. 2021. Automated vs. human health coaching: Exploring participant and practitioner experiences. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 99:1–99:37. <https://doi.org/10.1145/3449173>
- [44] Frode Moen and Einar Skaalvik. 2009. The effect from executive coaching on performance psychology. *International Journal of Evidence Based Coaching & Mentoring* 7, 2 (2009), 31–49.
- [45] Ine Mols, Elise van den Hoven, and Berry Eggen. 2016. Informing design for reflection. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*. ACM, New York, NY, 1–10. <https://doi.org/10.1145/2971485.2971494>
- [46] Jase Moussa-Inaty. 2015. Reflective Writing through the Use of Guiding Questions. *International Journal of Teaching and Learning in Higher Education* 27, 1 (2015), 104–113.
- [47] Jaya Narain, Tina Quach, Monique Davey, Hae Won Park, Cynthia Breazeal, and Rosalind Picard. 2020. Promoting wellbeing with sunny, a chatbot that facilitates positive messages within social groups. In *Extended Abstracts of the 2020 ACM CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–8. <https://doi.org/10.1145/3334480.3383062>
- [48] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th Annual Conference on Neural Information Processing Systems*. Curran Associates, Inc., Red Hook, NY, 27730–27744.
- [49] Jonathan Passmore and David Tee. 2023. Can chatbots replace human coaches? Issues and dilemmas for the coaching profession, coaching clients and for organisations. *The Coaching Psychologist* 19, 1 (2023), 47–54. <https://doi.org/10.53841/bpstep.2023.19.1.47>
- [50] Xiang-Zhi Qiu, Tina Chien-Wen Yuan, Nanyi Bi, Ming-Chyi Huang, and Chuang-Wen You. 2023. Exploring the challenges and opportunities in developing systems to improve alcohol use disorder through chatbot technology. In *Extended Abstracts of the 2023 ACM CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 123:1–123:5. <https://doi.org/10.1145/3544549.3585635>
- [51] Heleen Rutjes, Martijn C. Willemsen, and Wijnand A. IJsselstein. 2019. Beyond behavior: The coach's perspective on technology in health coaching. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 670. <https://doi.org/10.1145/3290605.3300900>
- [52] Kathleen Ryan, Samantha Dockray, and Conor Linehan. 2022. Understanding how eHealth coaches tailor support for weight loss: Towards the design of person-centered coaching systems. In *Proceedings of the 2022 ACM CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 285:1–285:16. <https://doi.org/10.1145/3491102.3501864>
- [53] Malak Sadek, Rafael A. Calvo, and Céline Mougenot. 2023. Trends, Challenges and Processes in Conversational Agent Design: Exploring Practitioners' Views through Semi-Structured Interviews. In *Proceedings of the 5th International Conference on Conversational User Interfaces*. ACM, 13:1–13:10. <https://doi.org/10.1145/3571884.3597143>
- [54] Samiha Samrose, Daniel McDuff, Robert Sim, Jina Suh, Kael Rowan, Javier Hernandez, Sean Rintel, Kevin Moynihan, and Mary Czerwinski. 2021. MeetingCoach: An Intelligent Dashboard for Supporting Effective & Inclusive Meetings. In *Proceedings of the 2021 ACM CHI Conference on Human Factors in Computing Systems*. ACM, 252:1–252:13. <https://doi.org/10.1145/3411764.3445615>
- [55] Anselm L. Strauss and Juliet M. Corbin. 1990. *Basics of qualitative research: Grounded theory procedures and techniques*. Sage Publications, Newbury Park, CA.
- [56] Ingo Susing, Suzy Green, and Anthony M Grant. 2011. The potential use of the authenticity scale as an outcome measure in executive coaching. *The Coaching Psychologist* 7, 1 (2011), 16–25.
- [57] Nicky Terblanche and Danie Cilliers. 2020. Factors that influence users' adoption of being coached by an artificial intelligence coach. *Philosophy of Coaching: An International Journal* 5, 1 (2020), 61–70. <https://doi.org/10.22316/poc/05.1.06>
- [58] Nicky Terblanche, Joanna Moly, Kevin Williams, and Jeanette Maritz. 2022. Performance matters: Students' perceptions of artificial intelligence coach adoption factors. *Coaching: An International Journal of Theory, Research and Practice* 16, 1 (2022), 100–114. <https://doi.org/10.1080/17521882.2022.2094278>
- [59] Viswanath Venkatesh, Michael G. Morris, Gordon B. Davis, and Fred D. Davis. 2003. User acceptance of information technology: Toward a unified view. *MIS Quarterly* 27, 3 (2003), 425–478. <https://doi.org/10.2307/30036540>

- [60] Nadine Wagener, Leon Reicherts, Nima Zargham, Natalia Bartłomiejczyk, Ava Elizabeth Scott, Katherine Wang, Marit Bentvelzen, Evropi Stefanidi, Thomas Mildner, Yvonne Rogers, and Jasmin Niess. 2023. SelVReflect: A guided VR experience fostering reflection on personal challenges. In *Proceedings of the 2023 ACM CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 323:1–323:17. <https://doi.org/10.1145/3544548.3580763>
- [61] Philip Weber, Faisal Mahmood, Michael Ahmadi, Vanessa von Jan, Thomas Ludwig, and Rainer Wieching. 2023. Fridolin: Participatory Design and Evaluation of a Nutrition Chatbot for Older Adults. *i-com* 22, 1 (2023), 33–52. <https://doi.org/10.1515/ICOM-2022-0042>
- [62] Thure Georg Weimann, Hannes Schlieter, and Alfred Benedikt Brendel. 2022. Virtual coaches: Background, theories, and future research directions. *Business & Information Systems Engineering* 64, 4 (2022), 515–528.
- [63] Laura Whitworth, Karen Kimsey-House, Henry Kimsey-House, and Phillip Sandahl. 1998. *Co-active coaching: New skills for coaching people toward success in work and life*. Nicholas Brealey, Boston, MA.
- [64] Robert Witherspoon and Randall P White. 1996. Executive coaching: A continuum of roles. *Consulting Psychology Journal: Practice and Research* 48, 2 (1996), 124–133. <https://doi.org/10.1037/1061-4087.48.2.124>
- [65] Wendy Wlasak, Sander Paul Zwanenburg, and Chris Paton. 2023. Supporting autonomous motivation for physical activity with chatbots during the COVID-19 pandemic: Factorial experiment. *JMIR Formative Research* 7 (2023), e38500. <https://doi.org/10.2196/38500>
- [66] Irmtraud Wolfbauer, Viktoria Pammer-Schindler, Katharina Maitz, and Carolyn P. Rosé. 2022. A Script for Conversational Reflection Guidance: A Field Study on Developing Reflection Competence With Apprentices. *IEEE Transactions on Learning Technologies* 15, 5 (2022), 554–566. <https://doi.org/10.1109/TLT.2022.3207226>
- [67] Alex M. Wood, P. Alex Linley, John Maltby, Michael Baliousis, and Stephen Joseph. 2008. The authentic personality: A theoretical and empirical conceptualization and the development of the authenticity scale. *Journal of Counseling Psychology* 55, 3 (2008), 385–399. <https://doi.org/10.1037/0022-0167.55.3.385>
- [68] Anna Xygkou, Panote Siriaraya, Alexandra Covaci, Holly Gwen Prigerson, Robert Neimeyer, Chee Siang Ang, and Wan-Jou She. 2023. The "conversation" about loss: Understanding how chatbot technology was used in supporting people in grief. In *Proceedings of the 2023 ACM CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 646:1–646:15. <https://doi.org/10.1145/3544548.3581154>
- [69] Wenbing Zhao, Xiongyi Liu, Tie Qiu, and Xiong Luo. 2020. Virtual avatar-based life coaching for children with autism spectrum disorder. *Computer* 53, 2 (2020), 26–34. <https://doi.org/10.1109/MC.2019.2915979>