Please quote as: Göldi, A., Rietsche, R. & Ungar, L. (2025). Efficient Management of LLM-Based Coaching Agents' Reasoning While Maintaining Interaction Quality and Speed. Conference on Human Factors in Computing Systems (CHI), Yokohama, Japan.

Efficient Management of LLM-Based Coaching Agents' Reasoning While Maintaining Interaction Quality and Speed

Andreas Göldi
University of St.Gallen
St.Gallen, Switzerland
andreas.goeldi@unisg.ch

Roman Rietsche
Institute for Digital Technology Management
Bern, Bern, Switzerland
roman.rietsche@bfh.ch

Lyle Ungar
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, Pennsylvania, USA
ungar@cis.upenn.edu

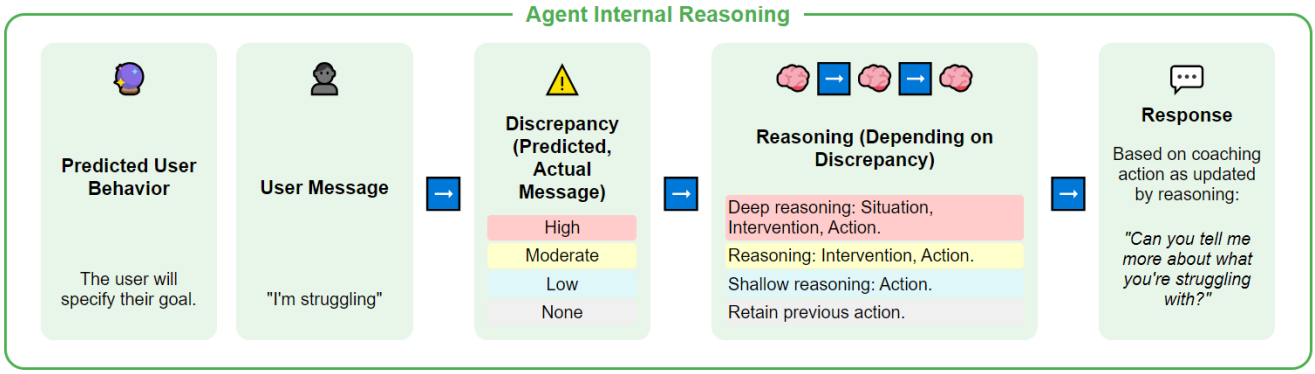


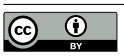
Figure 1: Illustrates how discrepancy can be used to manage reasoning depth in an LLM-based agent for coaching.

Abstract

LLM-based agents improve upon standalone LLMs, which are optimized for immediate intent-satisfaction, by allowing the pursuit of more extended objectives, such as helping users over the long term. To do so, LLM-based agents need to reason before responding. For complex tasks like personalized coaching, this reasoning can be informed by adding relevant information at key moments, shifting it in the desired direction. However, the pursuit of objectives beyond interaction quality may compromise this very quality. Moreover, as the depth and informativeness of reasoning increase, so do the number of tokens required, leading to higher latency and cost. This study investigates how an LLM-based coaching agent can adjust its reasoning depth using a discrepancy mechanism that signals how much reasoning effort to allocate based on how well the objective is being met. Our discrepancy-based mechanism constrains reasoning to better align with alternative objectives, reducing cost roughly tenfold while minimally impacting interaction quality.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; *HCI design and evaluation methods.*



CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713606

Keywords

Behavior Change, Education/Learning, Text/Speech/Language, Artifact or System, Quantitative Methods

ACM Reference Format:
Andreas Göldi, Roman Rietsche, and Lyle Ungar. 2025. Efficient Management of LLM-Based Coaching Agents' Reasoning While Maintaining Interaction Quality and Speed. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3706598.3713606

1 Introduction

Large Language Models (LLMs) have significantly impacted human-computer interaction; their integration into AI agents marks a step toward broader application versatility [26, 79]. LLM-based agents surpass standalone LLMs by internally generating chains [75] of text before responding to the end user. This enables more sophisticated informed reasoning processes [76]. Unlike standalone LLMs, which are optimized for immediate intent satisfaction through Reinforcement Learning from Human Feedback [14], LLM-based agents can incorporate extended objectives, such as helping users over the long term. For instance, by reasoning about how to stay aligned with developer-provided principles, they can respond more ethically [61]. Similarly, an informed reasoning process can align agent responses with theory-informed expertise. In this way, agents can respond more helpfully. To help users over the long term, a coaching agent, for instance, can use this approach to professionalize its dialogue, turning the surface-level conversation enabled by the

LLM itself to a helpful coaching, and support user goal pursuit over the long term.

To align agents with long-term objectives beyond immediate intent satisfaction, a structured agentic workflow or cognitive architecture can be implemented [60], where reasoning is deliberately guided toward the desired objectives at critical moments. This involves configuring the cognitive architecture to extract and process relevant information during intermediate reasoning steps. For example, in a coaching chatbot, the system could extract key elements from prior dialogues, such as identifying obstacles to a user's goals, and incorporate these insights into the reasoning sequence. Although earlier research on dialogue management focused on maintaining conversational consistency [9], LLMs have resolved mainly these issues [11]. However, many systems still emphasize short, single-session interactions, overlooking the need for long-term, personalized dialogue. To address this, reasoning mechanisms that maintain a consistent persona or use targeted summaries can enable more appropriate and effective long-term responses [44]. Furthermore, in complex dialogues, single-step inferences often fail to uncover hidden user needs, which requires intermediate reasoning steps [24, 70]. The current challenge lies in aligning reasoning processes with long-term objectives and designing informed reasoning processes tailored for LLM-based agents.

However, while informed reasoning can align agent responses with objectives beyond immediate intent satisfaction, it has potential downsides. First, introducing objectives beyond intent satisfaction can negatively impact the perceived quality of interactions. The former may suffer when balancing the competing goals of interaction quality and achieving an external objective. Second, incorporating additional information to structure reasoning can increase token generation, leading to higher costs, increased latency, and greater energy consumption. These drawbacks can potentially be mitigated by calibrating the informed reasoning process based on the extent of the discrepancy from the agent's external objective. For example, a coaching agent designed to help users set specific goals may require minimal reasoning if the user naturally provides clear, actionable goals. On the other hand, if a user consistently sets vague goals, the agent would need to engage in more extensive reasoning to guide them toward greater specificity.

Agents designed to pursue long-term objectives may generate responses that users like less than those from standard LLMs, which are typically optimized for immediate, short-term responses. This shift in focus can potentially undermine short-term interaction quality, as the agent prioritizes broader goals over immediate intent satisfaction. Furthermore, the complex reasoning required to align with external objectives may lead to higher token costs. Consequently, two key questions emerge:

RQ1: *How does aligning reasoning with external objectives affect interaction quality?*

RQ2: *How can reasoning be made cost-efficient while maintaining alignment with external objectives?*

To explore these questions, we investigated four versions of an LLM-based agent - a coaching chatbot designed to help users set and define their learning goals. Coaching was chosen as it provides a well-suited context for studying the effects of informed reasoning on user perceptions of LLM-based agents. Coaching, as a dialogic practice [1], aims to align a coachee's thoughts and behaviors with their motivations [55], and has been shown to be effective in promoting goal attainment for adult learners [15, 67]. While human coaches can achieve these outcomes, they are often too expensive for widespread use, underscoring the need for AI-driven coaching solutions [56, 64]. Since coaching requires longer-term interactions to produce meaningful improvements in coachees' lives, so standalone LLMs are insufficient. However, LLM-based agents incorporating evidence-based reasoning into their responses present a promising alternative.

The four versions of the coaching chatbot are as follows. First, we start with a baseline (1) naive agent that lacks additional reasoning processes. This agent is compared to versions with (2) a single level of informed reasoning, (3) two levels of informed reasoning without a discrepancy-based mechanism, and (4) a comprehensive agent that integrates two levels of informed reasoning alongside a discrepancy-based mechanism to optimize reasoning effort. To assess whether these different constraints on the reasoning process affect interaction quality, we measured variables such as expectation confirmation, interaction enjoyment, and satisfaction in a randomized experiment (n=86). To evaluate cost efficiency, we tracked token usage throughout the interaction. Our study makes two main contributions:

First, we demonstrate that balancing cost efficiency with interaction quality is achievable by implementing a discrepancy mechanism. This mechanism effectively reduces token generation, latency, and energy use while maintaining interaction quality. Dynamically adjusting reasoning based on user input ensures that resources are utilized efficiently without compromising the user experience.

Second, we validate this approach within a coaching context. We show that interaction quality remains high even when the agent prompts itself to aim for a goal beyond immediate intent satisfaction–specifically, coaching success. While our results do not confirm that long-term objectives are fulfilled, we demonstrate that striving toward them does not undermine short-term interaction quality. This distinction between short-term and long-term objectives is crucial: short-term objectives, such as immediate interaction quality, represent what users immediately perceive during the interaction. These are essential for meeting user expectations and ensuring engagement. Long-term objectives, such as the sustained goal attainment of the coachee, focus on outcomes that extend beyond the immediate interaction and are central to achieving meaningful change. Our findings establish that prioritizing long-term goals can coexist with maintaining a high standard of short-term interaction quality. This balance is a necessary prerequisite for future research and practice to focus on understanding and optimizing long-term effects.

These contributions form essential building blocks for future investigations into LLM-based agents that aim for long-term objectives without compromising immediate user satisfaction. Short-term objectives, such as maintaining immediate interaction quality, are critical to meeting user expectations but may come at the expense of long-term goals. In coaching, multi-step reasoning informed by coaching theory allows agents to guide users more effectively toward sustained goal achievement. However, it remains unclear whether aligning reasoning with external goals impacts interaction quality.

2 Background and Related Work

LLM-based agents must balance immediate interaction quality with broader long-term objectives. One key approach to managing this balance is using discrepancy mechanisms, which dynamically regulate reasoning depth. To understand the role of these mechanisms, we first explore their foundational concept in dialogue and then examine their application in coaching interactions.

2.1 Comparison with Related Work

Our research contributes to the field of LLM-based agents by addressing the trade-off between reasoning depth, interaction quality, and computational cost in personalized coaching. While recent advancements in coaching chatbots [32, 59] focus on task-specific effectiveness and hybrid human-AI collaboration, they lack mechanisms to dynamically adjust reasoning depth based on the fulfillment of objectives. Similarly, dialogue state tracking research [12, 40] emphasizes managing topic diversity and cross-domain adaptability but does not address the cost-quality balance central to our work. Studies on LLM-based AI agents [10, 21] broadly explore agent architectures and control mechanisms without incorporating discrepancy-based mechanisms to optimize reasoning effort. Furthermore, research on balancing performance and cost [7, 74] highlights heuristic reasoning and adaptive learning. Still, these approaches do not align with the personalized and cost-efficient reasoning adjustments demonstrated in our work. Our discrepancy-based mechanism thus fills an important gap by enabling LLM-based agents to achieve long-term objectives effectively while significantly reducing computational overhead.

2.2 Discrepancy as a Signal in Dialogue

Discrepancy is the difference between current behavior and the intended objective, such as in a goal-oriented conversation. Discrepancies between current performance and desired outcomes drive humans to adjust their actions to close this gap [57]. This process is essential for achieving long-term objectives, as discrepancies between realities and goals act as a motivating force, prompting individuals to reassess their strategies and align them with their objectives [16]. For example, suppose a client's progress toward a personal goal (e.g., passing an exam) falls short of expectations in a coaching interaction. In that case, the coach might identify the discrepancy and suggest adjustments, such as modifying habits, to better align with the goal. However, unresolved discrepancies, such as goal conflicts, can negatively impact well-being, whereas aligning goals with intrinsic motivations enhances well-being [35]. Goal conflicts divide resources, as those allocated to pursuing one goal cannot be used for another [47]. Therefore, concurrently pursuing multiple objectives entails potential downsides.

Discrepancy is critical in modulating reasoning depth and balancing interactions in dialogue management. Discrepancy mechanisms adjust the amount of reasoning effort an agent expends based on the alignment between observed user behavior and the conversation's goals [28, 58]. By signaling when further reasoning is required, these mechanisms enable more efficient progress toward objectives while minimizing unnecessary complexity. However, while reducing discrepancies can help achieve goals more efficiently, it may also feel constraining to users, forcing them to adapt to the agent's guiding structure rather than engaging in a more free-flowing, self-determined conversation. Finding the right balance between achieving efficiency and maintaining user satisfaction is critical in goal-driven interactions, especially when long-term objectives are involved [36].

Next, we examine coaching, which considers the coachee's goals and aims to support them in their pursuit.

2.3 Coaching

Coaching is a dialogic practice [1] focused on helping individuals achieve personal and professional goals through personalized, goal-oriented conversations [55]. Psychological coaching for adult learners, particularly those pursuing educational or career objectives, aims to foster behavioral change, enhance resilience, and promote sustained achievement [5]. Coaching dialogues can take various forms–solution-oriented, dialogic, or dialectical–each providing a distinct pathway for coachees to reflect, clarify their goals, and make meaningful progress [51].

However, challenges arise when the coach's personality and behavior create barriers to success, as observed in human-to-human coaching settings [8]. While professional coaches are expected to demonstrate core competencies, many lack formal psychological training, which limits their ability to facilitate deeper behavioral change [48]. To address this issue, evidence-based coaching has emerged, emphasizing using empirical knowledge and structured methodologies to achieve more reliable outcomes [20]. This approach aligns well with psychological coaching, one of the four major coaching discourses (alongside managerial, network, and soul-guide coaching), which focuses on promoting behavior change [73]. Since professional coaches are highly educated and skilled, coaching is often prohibitively expensive at scale. Coaching chatbots aim to provide cost-efficient and accessible alternatives that, while not matching human coaching in quality, can offer scalable solutions [46]. Despite their potential, current coaching chatbots still lack the advanced reasoning capabilities necessary for facilitating long-term behavior modification, as highlighted in the research summary in Table 1.

Having established the general principles of coaching, we now explore how coaching sessions can be structured, focusing on the popular GROW model, which guides coachees toward achieving their goals in a systematic manner.

2.3.1 Coaching Session Structure. The GROW model (Goal, Reality, Options, Way Forward) remains one of the most widely used frameworks in psychological coaching, offering a step-by-step structure to align coaching sessions with goal achievement [20]. The process begins with defining the goal–what the coachee wants to achieve. Next, the coachee assesses the reality of their current situation, considering both challenges and available resources. This leads to the generation of multiple options, or potential strategies, to overcome obstacles. Finally, the session concludes by identifying actionable steps, known as the way forward, ensuring the coachee leaves with a clear plan to reach their goals. This structured approach is particularly beneficial for adult learners pursuing objectives such as obtaining professional certifications or improving career-related skills. Figure 2 illustrates the GROW model, showcasing how it provides a coherent and progressive coaching framework.

Table 1: Key Themes and Findings in Coaching Chatbot Research

Theme	Example Finding
User Engagement and Retention	Engagement improves with tailored features, such as personalized notifications, but sustained engagement remains challenging [77].
Behavioral Change and Health Coaching	Chatbots are effective in initiating short-term behavior change [68].
Personalization and Adaptation	Personalized coaching tailored to user preferences enhances effectiveness [69].
AI and Machine Learning Techniques in Coaching	AI techniques, such as transfer learning, improve dialogue flow and personalization [80].
Multimodal Interaction and Communication Modalities	Different communication methods like voice and text influence user satisfaction [66].
Accessibility and Inclusivity	Chatbots have shown potential in serving underserved populations [71].
Human-AI Hybrid Coaching Models	Hybrid models that combine human oversight with AI assistance support short-term goal tracking [65].
Efficacy of AI Coaches vs. Human Coaches	AI coaches achieve short-term goals, such as improving self-management behaviors [63].
Cultural and Linguistic Considerations	Multilingual and culturally tailored chatbots promise to increase engagement [80].

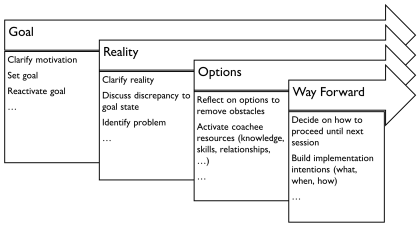


Figure 2: The GROW model structures a coaching session temporally, with different content depending on how far a conversation has reached. The conversation starts with a focus on the coachee's goal. Goals that our participants pursue are educational goals and include (random selection): "Public speaking certification", "I would like to complement my chemical engineering degree on this topic.", "Financial training.", "I'm thinking of doing a second degree to compliment the first one I got which was analytical chemistry.", "become a commercial pilot".

Finally, we turn to the role of empirically supported strategies for behavior change, which transform coaching dialogues from mere conversations into interventions that can help coachees pursue their goals more effectively.

2.3.2 Behavioral Change Techniques. Behavioral Change Techniques (BCTs) are a set of empirically validated strategies designed to facilitate behavior change [50]. These strategies may include breaking down tasks into manageable steps, offering rewards for progress, and reinforcing positive behaviors to encourage the coachee to adopt more effective habits. In psychological coaching, particularly in educational contexts, BCTs guide coachees from intention to action by systematically addressing barriers to change.

In coaching chatbots, BCTs are essential for moving beyond simple conversations to interventions that promote real behavior change [4]. When incorporated into chatbot architecture, these techniques provide the necessary structure to assist users in setting and achieving long-term goals. Because much of behavior change is communicated through instructions, feedback, and reinforcement, BCTs are highly applicable in AI-driven coaching systems [23]. By leveraging an established taxonomy of BCTs, coaching chatbots can deliver interventions that are not only personalized but also grounded in behavioral science, enhancing their effectiveness in promoting sustained change.

In summary, psychological coaching for adult learners requires structured reasoning to achieve immediate goals and long-term behavior change. This process relies on several key components: the reasoning depth needed to align with the coachee's evolving objectives, the use of temporal expertise such as the GROW model to structure sessions for progressive development, and the application of BCTs to guide interventions that promote actionable change. Additionally, discrepancy management plays a critical role by dynamically adjusting reasoning efforts based on the alignment between user behavior and coaching goals. While existing coaching chatbots are effective for short-term engagement, they often lack the reasoning depth and structured approach necessary to support sustained behavior change over time, as summarized in Table 1. By integrating structured reasoning, temporal expertise, BCTs, and discrepancy mechanisms, coaching chatbots can be significantly improved to better meet the demands of meaningful coaching conversations.

2.4 From Rule-based Chatbots to Agents

Historical and recent chatbot systems struggle with reasoning for use cases such as coaching. While traditional chatbots can handle simple interactions, they lack the depth needed to guide users toward objectives other than the immediate interaction itself.

2.4.1 Pre-LLM Chatbots. Chatbot technology has evolved significantly since its inception, beginning with foundational systems like ELIZA, which relied on fixed, rule-based approaches to simulate conversation [72]. While these early systems were groundbreaking, their rigidity and inability to handle complex user interactions often resulted in user dissatisfaction [45]. As conversational demands grew more sophisticated, chatbot technology advanced with the development of natural language understanding (NLU) systems. NLU chatbots improved upon their predecessors by interpreting more nuanced language inputs and managing user interactions more effectively. However, despite these advancements, they continued to face challenges, particularly with conversational breakdowns that could disrupt the flow of dialogue [2].

2.4.2 LLM-based Chatbots. The advent of LLMs brought about a significant leap in chatbot capabilities. LLMs, which generate responses based on vast datasets, enable more contextually relevant and adaptable interactions [25, 43]. These generative models have allowed chatbots to handle a wider range of queries and engage in more dynamic conversations. LLMs have been applied to coaching-adjacent domains, such as promoting well-being [37], or addressing social issues like reducing mental health stigma, where personalized interactions are crucial [42].

However, despite their strengths, LLM-based systems face notable challenges. Privacy concerns, issues related to prompt engineering, and the difficulty non-expert users encounter when interacting with these systems [78] remain significant obstacles to broader adoption [31]. Furthermore, coaching-specific applications of LLMs require a more structured dialogue approach that emphasizes aligning chatbot responses with the coachee's goals. In coaching contexts, mapping coachee-provided content to relevant coaching responses is essential for delivering effective and goal-oriented interventions [3].

2.4.3 LLM-based Agents. While LLM-based chatbots have improved conversational quality, they primarily focus on generating immediate, contextually relevant responses. In contrast, LLM-based agents extend beyond this by incorporating reasoning processes that allow them to pursue long-term objectives across multiple steps [26, 76]. Unlike chatbots, which react moment-to-moment, LLM-based agents are capable of autonomous reasoning–determining when additional information or deeper analysis (in the form of intermediate text generation that prompts the next step) is required before responding [13]. This capability makes agents better suited for complex tasks, such as providing coaching or facilitating behavior change, where understanding and aligning with long-term goals is essential.

A defining characteristic of LLM-based agents is their ability to integrate external data sources to enhance responses and guide their reasoning [13, 54, 75]. This feature is particularly valuable in coaching contexts, where agents may need to access additional information to assist users in achieving specific objectives over time. However, the coaching use case presents unique challenges. Coaching requires deeper reasoning capabilities that even current LLM-based chatbots lack, as noted in Table 1. LLM-based agents need to incorporate informed reasoning processes to effectively support users in achieving long-term goals. For example, they could use the GROW model to structure their understanding of the current coaching situation or select and apply BCTs to deliver more impactful coaching. The current gap in integrating these advanced reasoning capabilities highlights the need for further development to make LLM-based agents suitable for coaching and other applications requiring dialogues that produce long-term outcomes, not just immediate user satisfaction. To address this, managing both the dialogue flow and the reasoning behind the coaching agent's responses is critical.

2.5 Dialogue Management

Dialogue management shares many similarities with cognitive architectures in the context of conversational agents. Both focus on guiding the agent's reasoning processes toward achieving specific long-term objectives rather than merely satisfying immediate user intents. A structured agentic workflow or cognitive architecture configures the system to extract and process relevant information at key moments, ensuring the dialogue stays aligned with the user's goals [60]. In a coaching context, this means identifying critical elements, such as obstacles to a user's goals, from prior conversations and integrating them into the reasoning process to deliver more tailored, goal-oriented responses. Earlier efforts in dialogue management concentrated on maintaining conversational consistency, a challenge that recent advancements in LLMs have largely resolved [9].

2.5.1 Dialogue Interaction Quality. Suboptimal dialogue in conversational agents can lead to user dissatisfaction, particularly in coaching applications where responses are expected to be both personalized and empathetic [17]. This is especially true in sensitive contexts, such as mental health or coaching, where poorly designed interactions may leave users feeling misunderstood or disengaged [34]. Consequently, balancing high interaction quality with pursuing long-term objectives is essential.

Interaction quality can be assessed using five key metrics: expectation confirmation, performance expectancy, interaction enjoyment, satisfaction, and human-machine perception.

Expectation confirmation measures the extent to which the system meets user expectations [6]. In LLM-based coaching agents, users often expect fast, responsive outputs similar to those of standalone LLMs. However, when these agents engage in deeper reasoning–taking longer to deliver responses that support long-term coaching goals–this can lead to a mismatch between user expectations and actual system behavior, potentially reducing satisfaction.

Performance expectancy refers to users' belief that the system will help them achieve their goals [62]. In coaching applications, users expect the system to support their personal or professional outcomes effectively. If deeper reasoning or slower responses undermine this expectation, users may perceive the system as underperforming.

Interaction enjoyment captures users' emotional responses to their interactions. As reasoning depth increases and responses become slower or more complex, users may find the interaction less enjoyable, even if the system ultimately provides higher-quality guidance [41]. A decline in enjoyment could negatively impact the likelihood of continued use.

Satisfaction extends beyond interaction enjoyment to encompass the overall experience with the agent [30]. While deeper reasoning can lead to more thoughtful and accurate responses, delays or added complexity during real-time interactions may diminish users' short-term satisfaction.

Human-machine perception measures how users perceive the agent's identity [22]. Suppose the agent's behavior deviates too far from typical LLM outputs in favor of complex reasoning. In that case, users may perceive it as less human-like [19], which could negatively influence their overall interaction experience.

In summary, maintaining high-quality dialogue interaction is critical to fostering user engagement, particularly in coaching applications where users expect tailored, empathetic responses to help them achieve their goals. Interactions risk falling short of user expectations if these factors are not carefully managed, leading to dissatisfaction and disengagement.

2.5.2 Dialogue Objective Beyond Interaction Quality. While interaction quality is vital, LLM-based agents must also pursue long-term objectives, such as helping users achieve sustained behavior change. Achieving these extended objectives requires deeper reasoning and more structured dialogue than simple, immediate interactions. However, it is crucial that this pursuit does not come at the expense of short-term interaction quality. Our study focuses on how a discrepancy mechanism can balance these needs, enabling agents to engage in deeper reasoning to achieve long-term goals while maintaining the immediate interaction quality users expect.

In coaching applications, for instance, users may initially seek quick, satisfying responses. However, reaching long-term goals such as improved resilience or sustained behavior change demands the agent engage in more deliberate, reflective interactions. This can lead to slower, more thoughtful responses that may be less immediately gratifying but are essential for aligning with long-term coaching objectives.

Thus, dialogue management's challenge lies in balancing the need for immediate, high-quality interactions with the system's long-term objectives. Like most standalone LLMs, systems that focus on delivering quick, satisfying responses may fail to facilitate meaningful progress toward the user's larger goals. Conversely, agents that engage in deep, extended reasoning may risk reducing short-term user engagement.

2.5.3 Balancing Interaction Quality and Extended Dialog Objectives. Pre-LLM agent architectures have used discrepancy mechanisms to balance immediate task performance with long-term goals [28, 58]. These mechanisms can similarly be adapted to LLM-based agents to manage the trade-off between short-term interaction quality and the pursuit of extended dialog objectives.

In an LLM-based coaching agent, a discrepancy arises when there is a gap between the expected outcome of a coaching action and the user's actual behavior, typically observed in their next response. For instance, if the agent prompts the user to specify a goal but the user does not follow through, a discrepancy occurs. Conversely, no discrepancy exists if the user complies with the recommendation, allowing the agent to avoid deeper reasoning and preserve short-term interaction efficiency.

When the agent detects that the user's behavior deviates from expectations, it can reduce reasoning efforts to improve response times, especially in cases of low user engagement. This efficiency is critical for scalable systems, where managing resources ensures smooth and responsive interactions across multiple users [64]. In coaching scenarios, for example, the agent may increase reasoning depth only when it observes significant deviations from expected behavior, such as repeated resistance to goal-setting prompts.

This ability to adjust dynamically allows LLM-based agents to maintain a balance between fluid, responsive conversations known from LLM interactions and the achievement of long-term outcomes like behavior change and goal attainment. By leveraging a discrepancy mechanism, these agents may offer both immediate conversational satisfaction and sustained progress toward broader objectives.

In conclusion, effective dialogue management in LLM-based agents requires balancing interaction quality with long-term objectives. Discrepancy-based reasoning allows agents to dynamically modulate interaction depth based on user engagement, ensuring responsiveness while guiding users toward meaningful and sustained outcomes over time.

3 Artifact

In this section, we describe the design and implementation of our LLM-based coaching agent, which supports users in their goal pursuit through structured conversations. The agent integrates domain-specific coaching theory [23], including the GROW model [20], BCTs [50], and discrepancy mechanisms [28], to emulate the guidance provided by a human coach. We implemented the agent using the LangGraph framework[1], which enables us to dynamically steer its reasoning process, ensuring alignment with professional coaching practices.

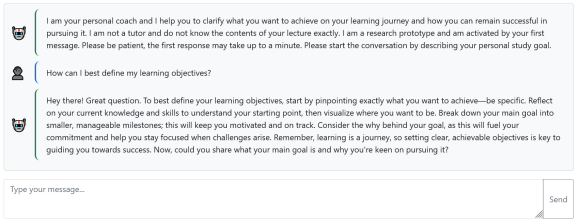


Figure 3: The UI for all groups.

3.1 Implementation of the LLM-based Agent

LLM-based agents rely on LLMs to generate text-based interactions that are responsive and adaptive to the user's input. These agents make multiple calls to the LLM, dynamically determining each next step. In our implementation, we use the LangGraph[2] framework, which facilitates chaining multiple LLM calls. This architecture allows us to incorporate domain-specific knowledge–such as coaching theories–during intermediate reasoning steps and to control the agent's behavior, ensuring alignment with professional coaching practices. This design ensures that each dialogue action is informed

[1] https://langchain-ai.github.io/langgraph/, version 0.2
[2] https://www.langchain.com/langgraph

by both the user's current state and the system's internal reasoning, enabling the agent to consistently apply structured coaching models like GROW and BCT throughout the conversation.

3.2 Aligning the Reasoning Process with Coaching Practice

To ensure the LLM agent's reasoning aligns with established coaching practices, we embed domain theory directly into the system's intermediate prompts [27]. The agent's cognitive architecture is designed to steer its responses toward meaningful coaching interventions, drawing on theories of human cognition [60]. This architecture enables the agent to provide structured, theory-based guidance that emulates professional coaching techniques.

3.2.1 Structuring the Coaching Conversation. The GROW model structures the coaching conversation into four stages: Goal, Reality, Options, and Way Forward [20]. We implement the GROW model in the agent by using function calling[3] to interpret the coachee's situation within these stages. This enables the coaching agent to determine what to focus on next. For example, if the goal is clearly specified, the agent can shift attention to understanding the coachee's current reality.

3.2.2 Behavioral Change Techniques: Goal-Oriented Interventions. In cases where the user struggles with procrastination–such as failing to meet study schedules despite wanting to pursue a certificate–the agent must go beyond merely offering encouragement. In these situations, behavioral interventions become essential. BCTs are evidence-based strategies designed to modify behavior, and we incorporate them into the agent's reasoning process [50]. For example, the agent might apply a BCT by prompting users to break their study goals into smaller, more manageable tasks.

We implement the selection of BCTs via function calls, enabling the agent to dynamically choose the most appropriate BCT based on the user's responses and their current stage in the GROW process. This approach ensures the agent stays focused on facilitating meaningful behavior change rather than engaging in off-topic dialogue.

3.2.3 Discrepancy Mechanism: Adjusting the Coaching Strategy. To maintain adaptability, our agent employs a discrepancy mechanism that evaluates the differences between the user's expected and actual responses during the session. Inspired by prior research on agent architectures [28, 58], this mechanism dynamically adjusts the agent's reasoning depth if the user's behavior deviates from the predicted path toward the overall objective. For instance, if the user is expected to brainstorm options but instead remains focused on their current realities, the agent can revise its understanding of the coaching situation and return to an earlier stage in the GROW model. Conversely, suppose the user follows the necessary steps to align with the agent's objective of helping them progress. No additional reasoning is required, and the agent can continue with its current approach.

[3]https://platform.openai.com/docs/guides/function-calling

3.3 LLM-based Agent Variants

To systematically explore the impact of varying reasoning depth on interaction quality and coaching effectiveness, we designed four variants of the coaching agent, each with varying levels of reasoning complexity. These variants were chosen to align with our study objectives: assessing how aligning reasoning with external objectives influences interaction quality (RQ1) and investigating how reasoning can be made cost-efficient while maintaining alignment with external objectives (RQ2). More specifically, we pit a (1) Naive agent with shallow reasoning on how to proceed against a (2) BCT agent that uses BCTs to inform this reasoning and against a (3) GROW+BCT agent which incorporates yet another reasoning level by analyzing the temporal situation of the coachee according to the GROW model. This third agent goes through all the reasoning levels each time while the (4) Full agent decides how deeply to reason based on the discrepancy, as do the (1) Naive and (2) BCT agents with their constrained repertoire of reasoning levels. By comparing these agents, we aim to uncover the trade-offs between immediate user satisfaction and the pursuit of long-term coaching goals (see the reasoning flow in Figure 4 and example dialogues, including the intermediate reasoning text generated to prompt the coaching messages, in Figure 5).

(1) The naive agent is the simplest version, reasoning only about the immediate next conversational action. It does not employ GROW, BCTs, or the discrepancy mechanism, focusing solely on maintaining dialogue with minimal reasoning complexity.
(2) The BCT-only agent focuses exclusively on behavioral guidance using BCTs, without employing GROW for session structure or the discrepancy mechanism for adjustments. This version provides actionable, behavior-focused advice but lacks the broader structure and adaptability of the more advanced agents.
(3) The GROW+BCT agent removes the discrepancy mechanism but retains the GROW and BCT reasoning layers. This version offers structured coaching through the GROW model and uses BCTs to guide the user's actions, though it does not recalibrate its strategy in response to user behavior deviations.
(4) The full agent integrates all three reasoning layers: GROW for session structuring, BCTs for goal-oriented interventions, and the discrepancy mechanism for dynamic adjustment. This version is the most comprehensive, offering both structured and adaptive coaching, guiding users through every stage of their goal pursuit.

3.4 Hypotheses and Expected Outcomes

Based on our agent designs, we propose the following hypotheses:

H1: The (4) Full agent, which integrates reasoning about the temporal stage (GROW model), behavioral interventions (BCTs), and dynamic adjustment (discrepancy mechanism), will be the most cost-efficient among the four agent variants.

H2: The added constraints on the (4) Full agent's reasoning may negatively impact immediate interaction quality compared to the (1) Naive and (2) BCT-only agents.

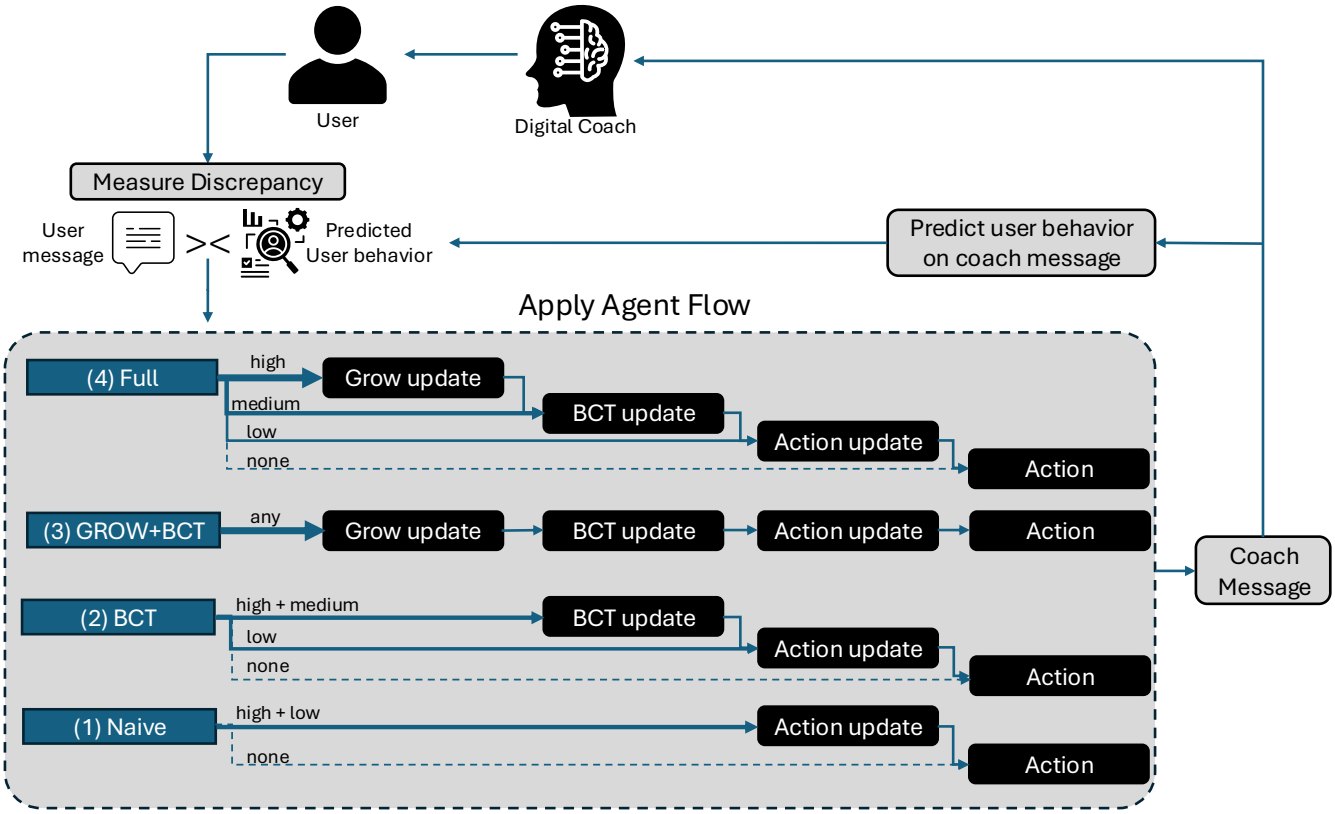


Figure 4: Flowchart of the reasoning process for the four agent variants (1) Naive, (2) BCT, (3) GROW+BCT, (4) Full. Each variant responds differently to discrepancies in dialogue when determining the next coaching action. (1) Naive agent: updates its action based on the full dialogue if a discrepancy is detected, otherwise retaining the previous action. (2) BCT agent: updates its Behavioral Change Technique (BCT) based on the full dialogue when there is a moderate discrepancy; for low discrepancies, it updates only the action based on the previous BCT. (3) GROW+BCT agent: updates the GROW model of the situation each time based on the full dialogue, followed by updates to the BCT and action, irrespective of discrepancies. (4) Full agent: Updates GROW understanding based on the full dialogue only for high discrepancies, updating BCT based on the GROW understanding for moderate discrepancies, and the action based on the prior BCT for low discrepancies. When no discrepancy is detected, no updates are made, optimizing efficiency. *Edge legend: Dotted lines indicate no discrepancy, faint lines indicate low discrepancy and bold lines indicate at least moderate discrepancy. Variant (3) GROW+BCT ignores discrepancy.*

H3: The (3) GROW+BCT agent, which engages in constant deep reasoning without dynamic adjustment, will be the least cost-efficient.

H4: The (1) Naive and (2) BCT-only agents will offer higher immediate interaction quality compared to the (4) Full agent.

These hypotheses are grounded in the expectation that the discrepancy mechanism in the (4) Full agent will optimize reasoning depth, reducing unnecessary token generation and resource usage (H1). However, it may detract from immediate interaction quality due to added reasoning constraints (H2). The (3) GROW+BCT agent, which lacks dynamic adjustment, is anticipated to be less cost-efficient (H3). Meanwhile, the simpler reasoning approaches of the (1) Naive and (2) BCT-only agents are expected to result in quicker responses and higher immediate interaction quality (H4).

The (3) GROW+BCT agent, engaging in heavy reasoning throughout interactions, is likely the least cost-efficient. It consistently applies deep reasoning without dynamic adjustment, leading to higher token usage and increased computational costs. Though structured, it lacks the efficiency benefits of discrepancy-based reasoning, resulting in suboptimal cost performance.

In contrast, the (1) Naive and (2) BCT-only agents, which more closely resemble standalone LLMs optimized for interaction quality [14], are expected to deliver higher immediate interaction quality. Their simpler reasoning processes allow for faster responses, aligning with user expectations. However, their limited ability to reason about long-term objectives may lead to the accumulation of unresolved discrepancies, requiring more adjustments over time.

Figure 5: The four versions with example dialogues, including reasoning, are in the yellow boxes. Reasoning is invisible to users. The initially longer reasoning of the (2) BCT and (4) Full agents becomes more efficient than that of the (1) Naive agent over time because more reasoning is triggered by more discrepancy, and their reasoning reduces discrepancy. (3) GROW+BCT Agent is less efficient since it does not use discrepancy but always uses reasoning to maximize depth.

While the (2) BCT-only agent incorporates some behavioral guidance through structured interventions, it demands less reasoning effort than the (3) GROW+BCT agent.

In summary, the (4) Full agent is expected to be the most cost-efficient due to its use of the discrepancy mechanism but may sacrifice short-term interaction quality. The (1) Naive and (2) BCT-only agents are anticipated to excel in immediate interaction quality but incur higher long-term costs. The (3) GROW+BCT agent, due to its constant deep reasoning, is likely to be the least cost-efficient of the four variants.

4 Methods

4.1 Participants

A total of 86 participants completed the study, engaging with the LLM-based agent and completing the post-experience survey. The final sample comprised 66% males and 34% females, with no participants identifying as other genders. The mean age of the participants was 31 years (SD = 9.0).

4.2 Procedure

Participants were recruited via Prolific.com[4] to take part in a between-subjects study. After providing consent, participants completed a brief demographic questionnaire. They were then randomly assigned to one of the four agent variants, which differed in their reasoning mechanisms (see Section 3 for details). After interacting with the coaching agent, participants completed a post-interaction survey measuring five dependent variables related to interaction quality and user experience. The entire process, including the chatbot interaction and post-interaction survey, took a median of 22 minutes. Participants were compensated at a standard rate of £2.20.

4.3 Operationalization of the Dependent Variables

We measured several dependent variables aligned with our study objectives to evaluate how our agent variants impact interaction quality and coaching effectiveness. Specifically, we focused on user satisfaction, interaction enjoyment, expectation confirmation, performance expectancy, and human-machine perception. These variables were chosen because they offer insights into users' perceptions of the agent's ability to meet their immediate needs while also supporting their long-term coaching goals. By assessing these factors, we aim to determine how different reasoning depths influence the overall effectiveness of the coaching interaction and directly address our research questions.

Expectation confirmation is the primary dependent variable and was measured using a Likert scale with three items [6]. This measure was selected based on the assumption that users typically expect immediate intent-satisfaction from a modern chatbot, as most well-known chatbots are optimized for this purpose [e.g., 52]. However, our agent's architecture incorporates additional reasoning levels to pursue extended objectives, such as supporting the user's long-term goal pursuit. This shift in focus may reduce immediate intent-satisfaction, potentially undermining user expectations and affecting confirmation. Expectation confirmation is also crucial because it predicts a range of other interaction quality metrics, such as satisfaction and continued use [49]. To ensure relevance, we adapted the measurement items to fit the experimental context. See the online Appendix A for details.

Performance expectancy [62] was measured because it is conceptually similar to expectation confirmation but focuses more specifically on users' beliefs about the chatbot's ability to meet performance-related expectations. While expectation confirmation examines the broader match between user expectations and outcomes, performance expectancy hones in on the anticipated functionality and usefulness of the chatbot agent. Including both measures provides a more nuanced understanding of user expectations, capturing different facets of the user experience.

Interaction enjoyment [41] was assessed to capture the emotional and affective aspects of the user experience. We hypothesized that chatbots constrained by more advanced reasoning processes, such as the (3) GROW+BCT variant, might deliver more structured but less enjoyable interactions, particularly if they deviate from the fluid and anthropomorphic behavior optimized by RLHF in standalone LLMs. Emotional engagement is a significant factor in determining overall satisfaction with technology and serves as an important predictor of continued use [38].

Satisfaction [29] was measured as a general measure that encompasses the overall intuitive evaluation of the chatbot interaction.

Human-machine perception [22] was included because we were particularly interested in how our agent variants might influence the perception of anthropomorphism [19]. LLMs optimized through RLHF are designed to enhance anthropomorphic qualities, but introducing more structured reasoning processes might negatively impact these perceptions. For instance, overly constrained reasoning may result in repetitive responses or other machine-like features that diminish the agent's human-like qualities.

In addition to these subjective measures, we also tracked each user's *token usage*, which correlates with latency, and the resulting *costs* for each interaction using LangSmith[5]. Furthermore, we calculated the relative costs per unit of interaction quality, as measured by the Likert scales discussed above (e.g., expectation confirmation). This analysis allowed us to evaluate not only the user experience but also the cost-effectiveness of each agent variant.

5 Results

5.1 Descriptive Statistics

Scale reliabilities ranged from α = 0.729 for the expectation confirmation measure to α = 0.933 for performance expectancy, indicating acceptable to excellent internal consistency across all scales. Significant correlations were observed between key variables, such as satisfaction and performance expectancy (r = 0.80) and human-machine perception and interaction enjoyment (r = 0.60). Correlations among outcome variables ranged from 0.60 to 0.80, demonstrating both convergent and discriminant validity. A detailed breakdown of reliability and interrelations, as well as means, can be found in the online Appendix (A). For example, expectation confirmation had a mean of 4.23 (SD = 0.97), while performance expectancy averaged 4.52 (SD = 0.87).

[4] prolific.com was chosen among online participant pool providers for its comparatively high response quality and ethical standards [53]

[5] smith.langchain.com

5.1.1 Primary Outcome: Effects on Expectation Confirmation. The primary outcome measure, indicating interaction quality and expectation confirmation, was subjected to both traditional null-hypothesis parametric tests[6] and equivalence testing. A one-way ANOVA revealed a significant effect of the agent condition on expectation confirmation, $F(3, 82) = 5.67$, $p = 0.004$. Post-hoc comparisons indicated that the full agent condition significantly outperformed the (1) Naive agent condition, adjusted $p = 0.006$, and the (2) BCT agent, adjusted $p = 0.028$, but not the(3) GROW+BCT agent. These results do not support Hypotheses H2 and H4, as the (4) Full agent did not show a decrease in immediate interaction quality compared to the (1) Naive and (2) BCT-only agents. Instead, the (4) Full agent performed as well as or better than the other agents on interaction quality measures. These differences were further analyzed using a Two One-Sided Test (TOST) to evaluate practical significance. The TOST procedure helps us check if two conditions are practically equivalent by setting upper and lower limits (called equivalence bounds) for what we consider a meaningful difference. If the observed difference falls within these bounds, we can confidently say the effect is too small to matter in practice. This test is important because it lets researchers statistically support the absence of a meaningful effect, rather than just concluding no effect based on nonsignificant results [39].

The TOST results confirmed that although statistically significant differences were observed, the effect sizes fell within predefined equivalence bounds, indicating that these differences, though present, are unlikely to be practically meaningful (see Table 2). Equivalence bounds were established using Cohen's d corresponding to an effect size with 80% power, multiplied by the pooled standard deviation. This approach ensures that only effect sizes exceeding a threshold of practical relevance fall outside the bounds. As the observed effect sizes did not surpass these bounds, the differences, while statistically significant, remain practically irrelevant. Therefore, we conclude that even though significant differences exist between the (2) BCT agent and the full agent (which uses GROW as an additional reasoning level and the discrepancy mechanism) that indicate a preference for the (2) BCT agent, this is within bounds.

5.1.2 Secondary Outcomes: Effects on the Four Additional Dependent Variables. The secondary outcomes included performance expectancy, interaction enjoyment, satisfaction, and human-machine perception. These were also analyzed using ANOVA and subsequent equivalence tests.

Performance expectancy demonstrated significant group differences, $F(3, 82) = 6.38$, $p = 0.002$. Similar to the expectation confirmation measure, post-hoc analysis revealed that the full agent significantly outperformed both the (1) Naive agent ($p < 0.01$) and the (2) BCT agent ($p < 0.05$). However, the TOST results confirmed that, despite statistical significance, these differences were practically small and within the predefined equivalence bounds.

ANOVA for *interaction enjoyment* showed no significant differences across the groups, $F(3, 82) = 1.52$, $p = 0.215$. The full agent had a marginally higher mean score compared to the other conditions, but these differences were not statistically or practically significant according to equivalence testing.

Satisfaction scores varied across conditions, with the full agent again outperforming the (1) Naive agent, $F(3, 82) = 4.25$, $p = 0.009$. Post-hoc analysis revealed significant differences between the full agent and the (1) Naive agent ($p = 0.01$), with equivalence testing again confirming the practical irrelevance of these differences.

Human-Machine Perception: This outcome showed moderate differences between conditions, $F(3, 82) = 3.75$, $p = 0.015$. Specifically, the full agent outperformed the (2) BCT agent ($p = 0.05$), but no significant differences were found between the (3) GROW+BCT agent and the other conditions.

5.2 Interaction Quality Relative to Cost

Cost-adjusted measures of interaction quality were analyzed using non-parametric tests due to violations of normality assumptions (see online Appendix: A). A Kruskal-Wallis test indicated significant group differences, $H(3) = 7.82$, $p = 0.02$, in cost-adjusted interaction quality. Post-hoc comparisons using Dunn's test revealed that the full agent significantly outperformed the (2) BCT agent ($Z = -2.34$, $p = 0.02$), and the (3) GROW+BCT agent also showed advantages over the (1) Naive agent. These findings support Hypotheses H1 and H3, confirming that the (4) Full agent was the most cost-efficient, and the (3) GROW+BCT agent was the least cost-efficient due to its constant deep reasoning without dynamic adjustment.

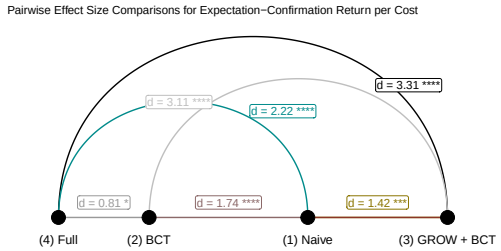


Figure 6: Illustrates how much more cost-efficient the (4) Full agent is compared to the other variants, on the representative example of expectation confirmation. $d = 2r/\sqrt{(1 - r^2)}$. *=p<**.05,** ***=p<**.001,** ***=p>**.0001. The (2) BCT variant is still much more cost-efficient than the remaining 2 variants because the discrepancy-mechanism only performs if reasoning depth can be adjusted (in the (1) Naive agent, discrepancy is assessed as a binary, meaning that there is no adaptive reasoning depth, only reasoning or no reasoning. (3) GROW+BCT has the most reasoning but does not perform as well as the costs of this would require for comparative efficiency to the other agent variants.** ***Performance, i.e.*** $\frac{\texttt{interaction quality}}{\texttt{cost}}$***, declines left to right.***

The data in Table 3 highlights the superior cost efficiency of the Full agent compared to other variants, with relative costs increasing up to 15x for the (3) GROW+BCT agent. The Full agent consistently achieved the lowest mean cents per user (0.00174) and the fewest tokens per user (8,487). Despite slightly fewer user messages, with a

[6]Distributions were normal, with slight ceiling and bottom effects still suitable for parametric testing [18].

Table 2: Ablation study, difference and equivalence tests for pairwise comparisons. We show raw and adjusted p values; p values were adjusted for multiple testing via Holm-correction. Difference tests test the Null-Hypothesis (NHST) that there are no differences. One difference is significant (bold). It shows that the (2) BCT agent outperforms the (4) Full agent in terms of expectation confirmation. However, all equivalence tests (two-sided tests or TOST) are significant, indicating that all values are within equivalence bounds and thus negligible in terms of practical significance.

Comparison		Difference	NHST p	NHST adj. p	TOST p	TOST adj. p
(4) Full vs	(1) Naive	-0.4848	0.0298	0.1383	3.53e-06	1.77e-05
(4) Full vs	(2) BCT	-0.5606	0.0062	**0.0369**	1.88e-07	1.13e-06
(4) Full vs	(3) GROW+BCT	-0.5197	0.0277	0.1383	5.80e-06	2.32e-05
(3) GROW+BCT vs	(1) Naive	-0.0758	0.6915	1.0000	3.54e-04	1.06e-03
(3) GROW+BCT vs	(2) BCT	-0.0348	0.8771	1.0000	2.84e-03	2.84e-03
(2) BCT vs	(1) Naive	0.0409	0.8412	1.0000	1.24e-03	2.49e-03

Table 3: Costs per agent variant. GROW+BCT without a discrepancy mechanism costs 15 times more than the full agent that deploys the same reasoning processes but only where necessary.

Group	Avg. Tokens/User	Cost Ratio (to Full)
(1) Naive	43,581	5x
(2) BCT	17,914	2x
(3) GROW+BCT	103,525	15x
(4) Full	8,487	1x

mean count of 8.32 (SD = 3.8), compared to the other agents, like (2) BCT (mean user messages = 10.41, SD = 5.13) and (3) GROW+BCT (m = 8.14, SD = 4.05), the (4) Full agent delivers the most cost-effective interaction quality. The naive agent was engaged with the least (m = 7.2, SD = 1.91). The Full agent, with its minimal cost and consistent performance, stands out as the most efficient choice for delivering high-quality interactions while keeping costs low. These findings underscore the practical and statistical significance of the Full agent's performance. For more statistical details, see the online Appendix (A).

5.3 Additional Qualitative Insights

We analyzed the participants' open text feedback on their experience to determine how users perceive the different variants of the coaching agent. To elicit this feedback, we asked *"In what way were your expectations using the digital coach confirmed?"* and *"In what way were your expectations using the digital coach NOT confirmed?"*. Answers were collected in the survey using an open answer format. We employed a thematic analysis on the overall feedback, also investigating differences between the groups that received (1) Naive, (2) BCT, (3) GROW+BCT, and (4) Full agents. We coded responses and grouped them into key **themes**[7]. This allowed us to examine how each variant impacted the user experience, particularly in terms of both confirmation and disconfirmation of expectations.

In terms of confirmation, four prominent themes emerged: **User-Friendliness**, **Adaptability and Guidance**, **Accuracy**, and **Artificiality Concerns**. Many participants emphasized the natural interaction with the digital coach, demonstrating high satisfaction with its user-friendliness. For example, one user from the (2) BCT group commented, *"It was a good experience, the coach was very useful and had good ideas."* Table 5 summarizes the key themes that were confirmed by participants, including definitions and examples. These themes highlight the strengths of the digital coach, particularly its user-friendliness and adaptability. The design implications offer actionable insights into how these strengths can be further enhanced in future iterations of the system.

However, disconfirmation arose when a participant added, *"It wasn't robotic or artificial seeming at all,"* suggesting that even though artificiality wasn't perceived as an issue, this group valued more human-like interaction. As shown in Table 6, participants also expressed concerns about the coach's response speed, artificiality, and the concreteness of its advice. For example, one participant noted, *"It took too long to get responses, which broke the flow of interaction,"* and another remarked, *"I expected more actionable guidance, but the advice felt too generic."* These comments underscore areas requiring improvement in responsiveness and personalization. These disconfirmation themes suggest areas where improvements could lead to better user satisfaction, such as refining response times and providing more personalized guidance. Still, many participants actually mentioned that their expectations were largely confirmed, e.g., *"I have not felt that my expectations were not confirmed"*.

For the (1) Naive variant, users appreciated the system's straightforwardness. A participant remarked, *"Complete answer and accurate."* However, some disconfirmation came from the system's lack of specificity, as another participant noted, *"The assistant wrote too much text and I think the half of it was innecessary."* This feedback suggests that while the system met basic expectations, users desired more tailored responses.

In the (3) GROW+BCT group, **Adaptability and Guidance** was highlighted as a strong theme. One participant confirmed, *"The conversation flowed and felt natural, and it gave some good solid advice, which is what I'd expect."* However, they disconfirmed their overall satisfaction by stating, *"The replies were slower than I was*

[7]We used GPT-4o to assist in the analysis. The results of all steps are logged in the codes_qual folder in the online Appendix A. The initial prompt was "I need a thematic analysis of these answers to two questions: In what way were your expectations using the digital coach confirmed? In what way were your expectations using the digital coach NOT confirmed? Step-by-step, following the appropriate guidelines". We provide the main conversation log as a text file (dialogicTA.txt) in the online Appendix A. Reporting is additionally based on manual checks and researcher discretion.

Table 4: Summary of Hypotheses and Corresponding Results

Hypothesis	Supported?	Evidence
H1: The (4) Full agent will be the most cost-efficient among the four agent variants.	Supported	The (4) Full agent had the lowest mean tokens per user and lowest cost, as shown in Table 3.
H2: The added constraints on the (4) Full agent's reasoning may negatively impact immediate interaction quality.	Not Supported	The (4) Full agent performed similarly or better on interaction quality measures, with differences within equivalence bounds.
H3: The (3) GROW+BCT agent will be the least cost-efficient.	Supported	The (3) GROW+BCT agent had the highest mean tokens per user and highest cost, as shown in Table 3.
H4: The (1) Naive and (2) BCT-only agents will offer higher immediate interaction quality.	Not Supported	The (4) Full agent did not show decreased interaction quality compared to the (1) Naive and (2) BCT-only agents.

Table 5: Themes from Users' Confirmed Expectations with Definitions, Examples, and Design Implications

Theme	Definition	Example	Design Implications
User-Friendliness	This theme captures the participants' positive experiences with how intuitive and easy the system was to use.	"The digital coach was really detailed and provided well-structured advice."	Ensure that the user interface is intuitive, reducing cognitive load and allowing for seamless interaction.
Adaptability & Guidance	This theme focuses on the digital coach's ability to provide personalized guidance and adapt to the user's requests.	"It adapted to my needs and provided helpful suggestions."	Focus on personalization, making sure the AI can adjust to different user needs and preferences dynamically.
Accuracy	This theme represents the participants' appreciation for the precise and accurate advice provided by the coach.	"The advice was on point and very accurate."	Increase precision in the information and suggestions provided by ensuring domain-specific expertise in responses.
Artificial Concerns	This theme addresses the participants' initial concerns about the AI seeming too robotic, which were alleviated during the interaction.	"Initially, I thought it would feel robotic, but it actually felt more human."	Humanize AI interactions by incorporating more natural language and empathetic responses to reduce robotic perceptions.

expecting, and the coach asked me quite a few questions on what I thought I should do, which I didn't really like." This indicates that while the GROW framework provided structure, some users found the method slower and overly reliant on self-reflection.

The (4) Full variant, which included all components, was generally well-received. One user confirmed, *"I expected and received intelligent and understanding communication,"* signaling satisfaction with the coach's adaptability. Nonetheless, disconfirmation was also present: *"There was no such thing. Everything was fine."* This suggests that while there were no major issues, users may have found the overall interaction somewhat uneventful. For instance, one participant stated, *"I didn't experience any problems with the system,"* indicating that for some users, the digital coach met their expectations satisfactorily without any notable concerns.

Table 7 highlights the specific experiences of users with each variant of the digital coach, showing which themes were present across the different conditions. It is clear that user-friendliness and adaptability were perceived more positively in certain variants, with the Full and *No GROW* variants showing the highest positive impact on these themes. Table 8 outlines the likely impact of each variant on key themes based on user feedback. This comparison indicates that while all variants have strengths, the Full variant likely provides the most balanced user experience, with high user-friendliness, adaptability, and accuracy while minimizing artificiality concerns.

In summary, different variants of the digital coach elicited diverse user reactions. (1) Naive participants found it accurate but desired more specificity. (2) BCT participants confirmed the coach's user-friendliness but sought a more human-like interaction. (3) GROW+BCT users valued the structured guidance but experienced slower interactions. Finally, the (4) Full variant achieved a balance, offering both adaptability and guidance with minimal disconfirmation.

Table 6: Themes from Users Disconfirmed Expectations with Definitions, Examples, and Design Implications

Theme	Definition	Example	Design Implications
Artificiality Concerns	This theme captures the residual concerns that the digital coach still felt somewhat robotic.	"It still felt a bit too robotic, lacking empathy."	Develop more empathetic and conversational dialogue strategies to minimize the robotic feel of interactions.
Response Speed	This theme highlights dissatisfaction with the time it took for the coach to respond.	"I felt frustrated by how long it took to get responses."	Optimize system response times to maintain the flow of conversation and reduce user frustration.
Concrete Help	This theme represents the participants' disappointment in the lack of specific, actionable advice from the coach.	"I expected more specific advice, but it was too generic."	Focus on providing clear, actionable steps and personalized recommendations to enhance the usefulness of interactions.
No Issues	This theme covers the responses where participants had no particular concerns or negative feedback.	"There were no significant issues or concerns."	Continue to ensure that the system meets basic expectations without introducing unnecessary complexity.
User-Friendliness	This theme acknowledges the participants' surprise at how natural and intuitive the interaction felt, despite initial concerns.	"I was surprised by how natural the interaction was, though I expected it to be worse."	Maintain intuitive design and improve onboarding processes to set proper expectations for first-time users.

Table 7: Impact of Agent Variants on Key Themes

Theme	(1) Naive	(2) BCT	(3) GROW+BCT	(4) Full
User-Friendliness	"The conversation flowed and felt natural."	"The responses I got for my questions were accurately worded."	"The conversation flowed and felt natural."	"It didn't seem robotic at all."
Adaptability and Guidance	N/A	"The digital coach adapted very well to my requests."	"I wanted to get some guidance from the digital coach."	"The coach adapted very well to my situation."
Accuracy	"The conversation felt natural, but I was expecting more precise advice."	"Complete answer and accurate."	"The coach gave me guidance that was precise."	"Responses were accurate and relevant."
Artificiality Concerns	"The interaction felt natural and not artificial."	N/A	"It still felt robotic at times."	"It felt more human and caring."

6 Discussion

This study evaluated four approaches to managing the reasoning depth of LLM-based agents by testing different versions of a coaching agent. LLM-based agents rely on LLMs for multi-step reasoning, iteratively prompting themselves before delivering a final response. The main feature under investigation was a discrepancy mechanism that dynamically adjusts reasoning effort based on how well user input aligns with the agent's objectives. The results showed that incorporating a discrepancy mechanism reduced token usage by more than tenfold while largely preserving interaction quality. Specifically, agent variants equipped with this mechanism enabled more cost-efficient interactions compared to the (3) GROW+BCT agent, which consistently applied the most reasoning effort using both the GROW coaching framework [20] and BCTs [50]. Additionally, while the (2) BCT-only agent achieved the highest short-term interaction quality, it incurred increased computational costs. This study thus supports the hypothesis that discrepancy can be used to efficiently manage LLM-based agents' reasoning depth while maintaining interaction quality and speed.

This study makes two primary contributions. First, it demonstrates that balancing cost-efficiency with interaction quality is

achievable by implementing a discrepancy mechanism in LLM-based agents. By dynamically adjusting reasoning depth based on user input, the discrepancy mechanism reduces unnecessary token generation, thereby improving latency and cost-efficiency without compromising interaction quality. This addresses the first research question, showing that a dynamic approach to managing reasoning depth can optimize the efficiency of LLM-based agents while maintaining user satisfaction.

Second, we validate this approach within a coaching context, demonstrating that interaction quality is maintained even when the agent pursues a goal beyond immediate intent satisfaction (namely, coaching success). Our findings suggest that agents can be designed to pursue long-term objectives without compromising short-term interaction quality. This is a necessary precondition for research and practice to focus on understanding long-term effects. Low short-term interaction quality would likely serve as a barrier to long-term success. While incorporating structured reasoning processes like GROW and BCTs could potentially result in more deliberate responses, our findings indicate that the impact on immediate user satisfaction is minimal.

Table 8: Likely Impact of Agent Variants on Themes

Theme	(1) Naive	(2) BCT	(3) GROW+BCT	(4) Full
User-Friendliness	High due to natural flow	Moderate due to mechanical feel	High but lacks concreteness	High, balanced user experience
Adaptability and Guidance	N/A, more free-form	Moderate, structured but less flexible	Moderate, slower response	High, full guidance present
Accuracy	Moderate, less structured	High, accurate responses	Moderate, accuracy hindered by slow responses	High, accurate and relevant
Artificiality Concerns	Low concerns, felt natural	Moderate concerns, slightly mechanical	Moderate concerns due to delay	Low concerns, more human-like
Response Speed	Unclear	Unclear	Likely slower, frustrating	Likely fast and efficient
Concrete Help	Unclear	Unclear	Likely lower, lacks specificity	Likely high, concrete advice present

Our findings indicate that varying the reasoning depth in LLM-based agents had minimal impact on user satisfaction and enjoyment. This suggests that users may tolerate increased reasoning complexity as long as the interaction remains engaging and responsive. According to Lee [41], while deeper reasoning can potentially reduce short-term interaction enjoyment due to longer response times, adaptive mechanisms can mitigate these effects. In our study, the use of discrepancy mechanisms likely helped maintain interaction quality by dynamically adjusting reasoning depth based on user input [28, 58]. This balance between thoughtful responses and user engagement suggests that agents can incorporate extended reasoning processes to pursue long-term objectives without significantly compromising user satisfaction. These implications are significant for designing LLM-based agents that need to maintain high interaction quality while achieving broader goals.

On the second question, regarding how reasoning for external objectives can be made cost-efficient, we observe that agents can achieve cost-efficiency if they employ deeper reasoning only when necessary, as regulated by a discrepancy mechanism. The (4) Full agent, which used the GROW model to gain a higher-level understanding of the coaching situation, required fewer updates than the BCT-only agent. This resulted in reduced token usage and fewer discrepancies over time, making the Full agent more cost-efficient for pursuing a long-term objective. Therefore, integrating additional reasoning levels, such as GROW, can be cost-efficient if paired with a discrepancy mechanism determining when the additional reasoning is beneficial.

This mirrors human cognition, which relies on two systems: one deliberate and one intuitive [33]. It may inform work on cognitive architectures for LLM-based agents [60], which aim to replicate human cognitive pathways in agents to produce more usable and predictable outputs. For human-computer interaction, such alignment could lead to the broader applicability of the vast corpus of psychological knowledge to understand users and the agents they interact with.

In examining the trade-offs between reasoning complexity and short-term interaction quality, we note that the (2) BCT-only agent generated stronger immediate expectation confirmation compared to the (4) Full agent, but at higher cost. This highlights that adding more structured reasoning, such as the GROW framework [20], does not always enhance short-term interaction quality. However, integrating GROW improves cost-efficiency because it operates with a broader understanding of the situation, requiring less frequent updates than BCT. The (4) Full agent demonstrated that incorporating additional reasoning layers slightly reduces short-term interaction quality, resulting in fewer discrepancies over time and ultimately lowering computational costs.

Further, we contribute practical design implications for developers of LLM-based coaching agents. The study demonstrates that agents combining the GROW model [20], BCTs [50], and a discrepancy mechanism can provide both high-quality user interactions and cost-effective operation. As Figure 4 illustrates, the (4) Full agent adjusts its reasoning steps based on the level of discrepancy detected, updating its understanding of the coaching situation only for high discrepancies while addressing lower levels of understanding, such as selecting the appropriate BCT, for moderate discrepancies. By integrating a discrepancy mechanism, developers can

ensure that LLM-based agents allocate reasoning effort only where necessary, reducing token generation costs for pursuing long-term objectives while maintaining interaction quality.

Reflecting on our findings, the choice of agent variants with varying reasoning depths–shallow, medium, and deep–was crucial in addressing our study goals: assessing how aligning reasoning with external objectives affects interaction quality (RQ1) and investigating how reasoning can be made cost-efficient while maintaining alignment with external objectives (RQ2). By systematically comparing agents with different reasoning complexities, we identified the trade-offs between immediate user satisfaction and the pursuit of long-term coaching goals. While our study focused on the coaching domain, the principles underlying the discrepancy-based mechanism have broader applicability. The coaching context provided an ideal testbed for these mechanisms. Still, similar strategies could be applied in other domains where agents need to balance immediate user satisfaction with long-term objectives. For instance, an AI assistant could leverage discrepancy mechanisms in legal advice systems to ensure it provides accurate and comprehensive legal information while maintaining an engaging and understandable interaction with the user. By integrating actual results into our discussion, we reinforce the connection between our methodological choices and the study goals, demonstrating that deeper reasoning processes can be both cost-efficient and effective in maintaining interaction quality while pursuing long-term objectives. This suggests that the discrepancy-based mechanism is a versatile design pattern that can be extended beyond coaching to enhance the performance of AI agents in diverse applications.

6.1 Limitations and Future Research

This study has several limitations. First, focusing on the coaching domain, specifically on learning goals such as certifications, may limit generalizability to other areas where the nature of objectives and user interactions differ. While we demonstrated the effectiveness of the discrepancy-based mechanism in a coaching context, its applicability to other domains remains to be tested. Future research should explore whether discrepancy-based reasoning can effectively balance short-term and long-term goals in customer service or financial planning contexts. In legal contexts, for instance, an AI assistant could employ discrepancy mechanisms to balance the provision of accurate legal counsel with user-friendly interactions, ensuring both compliance and accessibility. By examining the mechanism's effectiveness across various domains, it may be possible to identify common design patterns and enhance its adaptability, thereby contributing to broader AI agent development and offering insights into how agents can pursue extended objectives without undermining interaction quality. Second, we tested single-session interaction quality, not longer-term goal attainment, as the latter is of interest mainly to specialists studying coaching and coaching chatbots[8]. We aimed to inform the more preliminary yet broadly relevant question of whether it is possible to pursue outside objectives using LLM agents without compromising interaction quality and cost-efficiency. The third limitation relates to sample size and how it affects our interpretation of results. Null-hypothesis testing showed the BCT agent had better short-term interaction quality than the Full agent. Still, two-sided equivalence testing indicated this difference is within equivalence bounds, meaning it is not practically significant for most applications. However, in very large-scale real-world deployments, smaller effects may become statistically significant as equivalence bounds tighten with larger samples. Fourth, while this study focused on personal learning goals, the GROW model, and BCTs, results with other coaching methods or applications in non-coaching domains may differ. Similarly, the exclusive use of text-based interactions in this study leaves open the question of whether multimodal interfaces might alter the agent's effectiveness. Finally, this study did not address the broader implications of agents pursuing longer-term objectives that may deviate from immediate user intent. While this divergence aligns with coaching, where user goals take precedence, it may not be desirable in other domains where maintaining user intent is critical. For instance, in domains outside coaching, an agent's long-term reasoning could influence or manipulate users, raising ethical concerns. Despite these limitations, we hope our findings encourage investigations into LLM-based agents, as they demonstrate that the trade-off in interaction quality for agent objectives beyond intent satisfaction is minimal and that using a discrepancy-based mechanism can even turn more complex agent architectures more cost-efficient.

[8]We have measured goal attainment in long-term interaction with the coaching agent in a separate small sample (n=15 participated throughout an 8 week study period); results showed parity in goal attainment between students supported by humans or the agent. However, these results are preliminary and out of scope of the present study.

7 Conclusion

This study demonstrates that an LLM-based agent can pursue extended objectives with additional reasoning without compromising interaction quality or cost-efficiency. By using a discrepancy mechanism to adjust reasoning depth dynamically, agents can align long-term goals with short-term user expectations. This approach may be generalized across domains where agents must balance immediate interaction quality with broader objectives.

References

[1] Hilary Armstrong. 2012. Coaching as dialogue: Creating spaces for (mis) understandings. *International Journal of Evidence Based Coaching and Mentoring* 10 (2012), 33–47. Issue 1.

[2] Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D. Weisz. 2019. Resilient Chatbots. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300484

[3] Tatiana Bachkirova, Gordon Spence, and David B Drake (Eds.). 2017. *The SAGE handbook of coaching*. SAGE reference, Los Angles.

[4] Tessa Beinema, Harm op den Akker, Lex van Velsen, and Hermie Hermens. 2021. Tailoring coaching strategies to users' motivation in a multi-agent health coaching application. https://doi.org/10.1016/j.chb.2021.106787 Cited by: 19; All Open Access, Hybrid Gold Open Access.

[5] John L. Bennett and Francine Campon. 2017. *Coaching and Theories of Learning*. SAGE reference, Los Angles, Chapter 6, 102–121.

[6] Anol Bhattacherjee. 2001. Understanding Information Systems Continuance: An Expectation-Confirmation Model. *MIS Quarterly* 25 (9 2001), 351. Issue 3. https://doi.org/10.2307/3250921

[7] Robert Black and Jane Smith. 2024. How to Reduce Computation Time While Sparing Performance During Robot Navigation? A Neuro-Inspired Architecture for Autonomous Shifting Between Model-Based and Model-Free Learning. https://doi.org/10.48550/arXiv.2004.14698

[8] Anna Blackman, Alison Carter, and Rachel Hay. 2014. Coaching for effectiveness: initial findings from an international survey. https://api.semanticscholar.org/CorpusID:140755140

[9] Hayet Brabra, Marcos Baez, Boualem Benatallah, Walid Gaaloul, Sara Bouguelia, and Shayan Zamanirad. 2022. Dialogue Management in Conversational Systems: A Review of Approaches, Challenges, and Opportunities. *IEEE Transactions*

on Cognitive and Developmental Systems 14 (9 2022), 783–798. Issue 3. https://doi.org/10.1109/TCDS.2021.3086565

[10] Alice Brown and George White. 2024. Exploring Large Language Model Based Intelligent Agents: Definitions, Methods, and Prospects. https://doi.org/10.48550/arXiv.2401.03428

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[12] Emily Chen and Frank Zhao. 2024. Zero-shot Cross-domain Dialogue State Tracking via Context-aware Auto-prompting and Instruction-following Contrastive Decoding. https://doi.org/10.18653/v1/2024.emnlp-main.485

[13] Yulin Chen, Ning Ding, Hai-Tao Zheng, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Empowering Private Tutoring by Chaining Large Language Models. https://doi.org/10.48550/arXiv.2309.08112

[14] Paul F Christiano, Jan Leike, Tom B Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences.

[15] Elaine Cox. 2015. *Coaching and Adult Learning: Theory and Practice*. Jossey-Bass, San Francisco, California, 27–38.

[16] Charles B. Cross. 1997. The Modal Logic of Discrepancy. *Journal of Philosophical Logic* 26 (4 1997), 143–168. Issue 2. https://doi.org/10.1023/A:1017939604673

[17] Andrea Cuadra, Maria Wang, Lynn Andrea Stein, Malte F. Jung, Nicola Dell, Deborah Estrin, and James A. Landay. 2024. The Illusion of Empathy? Notes on Displays of Emotion in Human-Computer Interaction. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–18. https://doi.org/10.1145/3613904.3642336

[18] Todd A. DeWees, Gina L. Mazza, Michael A. Golafshar, and Amylou C. Dueck. 2020. Investigation Into the Effects of Using Normal Distribution Theory Methodology for Likert Scale Patient-Reported Outcome Data From Varying Underlying Distributions Including Floor/Ceiling Effects. *Value in Health* 23 (5 2020), 625–631. Issue 5. https://doi.org/10.1016/j.jval.2020.01.007

[19] Nicholas Epley, Adam Waytz, and John T. Cacioppo. 2007. On seeing human: A three-factor theory of anthropomorphism. *Psychological Review* 114 (2007), 864–886. Issue 4. https://doi.org/10.1037/0033-295X.114.4.864

[20] Anthony M. Grant. 2018. *Goals and coaching: an integrated evidence-based model of goal-focused coaching and coaching psychology*. Routledge, London, 34–50.

[21] Samantha Green and Robert Black. 2024. Formal-LLM: Integrating Formal Language and Natural Language for Controllable LLM-based Agents. https://doi.org/10.48550/arXiv.2402.00798

[22] Andreas Göldi. 2024. Flexibility in Chatbot Identity Perception. https://aisel.aisnet.org/amcis2024/ai_aa/ai_aa/3

[23] Andreas Göldi and Roman Rietsche. 2023. Whereto for Automated Coaching Conversation: Structured Intervention or Adaptive Generation? https://aisel.aisnet.org/ecis2023_rip/31

[24] Andreas Göldi and Roman Rietsche. 2024. Chatbot Agents Displaying Non-factive Reasoning Enhance Expectation Confirmation. https://aisel.aisnet.org/icis2024/humtechinter/humtechinter/8/

[25] Andreas Göldi and Roman Rietsche. 2024. Insert-expansions for Large Language Model Agents. https://aisel.aisnet.org/amcis2024/sig_hci/sig_hci/3

[26] Andreas Göldi and Roman Rietsche. 2024. Making Sense of Large Language Model-Based AI Agents. https://aisel.aisnet.org/icis2024/aiinbus/aiinbus/16/

[27] Andreas Göldi, Thiemo Wambsganss, Seyed Parsa Neshaei, and Roman Rietsche. 2024. Intelligent Support Engages Writers Through Relevant Cognitive Processes. , 12 pages. https://doi.org/10.1145/3613904.3642549

[28] Barbara Hayes-Roth. 1995. An architecture for adaptive intelligent systems. *Artificial Intelligence* 72 (1 1995), 329–365. Issue 1-2. https://doi.org/10.1016/0004-3702(94)00004-K

[29] SeJoon Hong, James Y.L. Thong, and Kar Yan Tam. 2006. Understanding continued information technology usage behavior: A comparison of three models in the context of mobile internet. *Decision Support Systems* 42 (12 2006), 1819–1834. Issue 3. https://doi.org/10.1016/j.dss.2006.03.009

[30] Qian Huang, Jian Vora, Percy Liang, and J Leskovec. 2023. Benchmarking Large Language Models As AI Research Agents. *ArXiv* abs/2310.03302 (2023), –. https://doi.org/10.48550/arXiv.2310.03302

[31] Eunkyung Jo, Yuin Jeong, Sohyun Park, Daniel A. Epstein, and Young-Ho Kim. 2024. Understanding the Impact of Long-Term Memory on Self-Disclosure with Large Language Model-Driven Chatbots for Public Health Intervention. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–21. https://doi.org/10.1145/3613904.3642420

[32] Peter Johnson and Mark Taylor. 2024. Coaching Copilot: Blended Form of an LLM-Powered Chatbot and a Human Coach to Effectively Support Self-Reflection for Leadership Growth. https://doi.org/10.48550/arXiv.2405.15250

[33] Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, NYC, NY, US. –499 pages.

[34] Rachael M. Kang and Tera L. Reynolds. 2024. "This app said I had severe depression, and now I don't know what to do": the unintentional harms of mental health applications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–17. https://doi.org/10.1145/3613904.3642178

[35] Rebecca E. Kelly, Warren Mansell, and Alex M. Wood. 2015. Goal conflict and well-being: A review and hierarchical model of goal conflict, ambivalence, self-discrepancy and self-concordance. *Personality and Individual Differences* 85 (10 2015), 212–229. https://doi.org/10.1016/j.paid.2015.05.011

[36] Mary C. Kernan and Robert G. Lord. 1990. Effects of valence, expectancies, and goal-performance discrepancies in single and multiple goal environments. *Journal of Applied Psychology* 75 (4 1990), 194–203. Issue 2. https://doi.org/10.1037/0021-9010.75.2.194

[37] Inyeop Kim and Uichin Lee. 2024. Navigating User-System Gaps: Understanding User-Interactions in User-Centric Context-Aware Systems for Digital Well-being Intervention. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–15. https://doi.org/10.1145/3613904.3641979

[38] Young Hoon Kim, Dan J. Kim, and Kathy Wachter. 2013. A study of mobile user engagement (MoEN): Engagement motivations, perceived value, satisfaction, and continued engagement intention. *Decision Support Systems* 56 (12 2013), 361–370. https://doi.org/10.1016/j.dss.2013.07.002

[39] Daniël Lakens. 2017. Equivalence Tests. *Social Psychological and Personality Science* 8 (5 2017), 355–362. Issue 4. https://doi.org/10.1177/1948550617697177

[40] Clara Lee and David Wang. 2024. S3-DST: Structured Open-Domain Dialogue Segmentation and State Tracking in the Era of LLMs. https://doi.org/10.18653/v1/2024.findings-acl.891

[41] SeoYoung Lee and Junho Choi. 2017. Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity. *International Journal of Human-Computer Studies* 103 (7 2017), 95–105. https://doi.org/10.1016/j.ijhcs.2017.02.005

[42] Yi-Chieh Lee, Yichao Cui, Jack Jamieson, Wayne Fu, and Naomi Yamashita. 2023. Exploring Effects of Chatbot-based Social Contact on Reducing Mental Illness Stigma. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–16. https://doi.org/10.1145/3544548.3581384

[43] Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. "I Hear You, I Feel You": Encouraging Deep Self-disclosure through a Chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376175

[44] Haoyuan Li, Hao Jiang, Tianke Zhang, Zhelun Yu, Aoxiong Yin, Hao Cheng, Siming Fu, Yuhao Zhang, and Wanggui He. 2023. TrainerAgent: Customizable and Efficient Model Training through LLM-Powered Multi-Agent System. *ArXiv* abs/2311.06622 (2023), –. https://doi.org/10.48550/arXiv.2311.06622

[45] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA". In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[46] Malte Lömker, Ulrike Weber, and Johannes Moskaliuk. 2021. *Chatbots im Coaching: Chancen im lösungs-fokussierten Coaching* (1st ed. ed.). Springer Fachmedien, Wiesbaden. https://doi.org/10.1007/978-3-658-32830-6

[47] Zita Mayer and Alexandra M. Freund. 2022. Better off without? Benefits and costs of resolving goal conflict through goal shelving and goal disengagement. *Motivation and Emotion* 46 (12 2022), 790–805. Issue 6. https://doi.org/10.1007/s11031-022-09966-x

[48] Ryan A. McKelley and Aaron B. Rochlen. 2007. The practice of coaching: Exploring alternatives to therapy for counseling-resistant men. *Psychology of Men & Masculinity* 8 (1 2007), 53–65. Issue 1. https://doi.org/10.1037/1524-9220.8.1.53

[49] Vicki McKinney, Kanghyun Yoon, and Fatemeh "Mariam" Zahedi. 2002. The Measurement of Web-Customer Satisfaction: An Expectation and Disconfirmation Approach. *Information Systems Research* 13 (9 2002), 296–315. Issue 3. https://doi.org/10.1287/isre.13.3.296.76

[50] Susan Michie, Michelle Richardson, Marie Johnston, Charles Abraham, Jill Francis, Wendy Hardeman, Martin P Eccles, James Cane, and Caroline E Wood. 2013. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. *Annals of behavioral medicine : a publication of the Society of Behavioral Medicine* 46 (2013), 81–95. Issue 1. https://doi.org/10.1007/s12160-013-9486-6

[51] Michael Neenan. 2008. From Cognitive Behaviour Therapy (CBT) to Cognitive Behaviour Coaching (CBC). *Journal of Rational-Emotive & Cognitive-Behavior Therapy* 26 (3 2008), 3–15. Issue 1. https://doi.org/10.1007/s10942-007-0073-2

[52] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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. ,

27730-27744 pages. https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf
[53] Stefan Palan and Christian Schitter. 2018. Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17 (3 2018), 22–27. https://doi.org/10.1016/j.jbef.2017.12.004
[54] Aaron Parisi, Yao Zhao, and Noah Fiedel. 2022. TALM: Tool Augmented Language Models. https://arxiv.org/abs/2205.12255v1
[55] Jonathan Passmore. 2018. *Behavioural Coaching*. Routledge, London, 99–107.
[56] 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 (6 2023), 47–54. Issue 1. https://doi.org/10.53841/bpstcp.2023.19.1.47
[57] Jean M. Phillips, John R. Hollenbeck, and Daniel R. Ilgen. 1996. Prevalence and prediction of positive discrepancy creation: Examining a discrepancy between two self-regulation theories. *Journal of Applied Psychology* 81 (10 1996), 498–511. Issue 5. https://doi.org/10.1037/0021-9010.81.5.498
[58] M. Scheutz and V. Andronache. 2004. Architectural Mechanisms for Dynamic Changes of Behavior Selection Strategies in Behavior-Based Systems. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)* 34 (12 2004), 2377–2395. Issue 6. https://doi.org/10.1109/TSMCB.2004.837309
[59] Jane Smith and Alice Brown. 2024. Unearthing AI Coaching Chatbots Capabilities for Professional Coaching: A Systematic Literature Review. https://doi.org/10.1108/jmd-06-2024-0182
[60] Theodore R Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L Griffiths. 2023. Cognitive Architectures for Language Agents. https://doi.org/10.48550/arXiv.2309.02427
[61] Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David D Cox, Yiming Yang, and Chuang Gan. 2023. Principle-Driven Self-Alignment of Language Models from Scratch with Minimal Human Supervision. *ArXiv* abs/2305.03047 (2023), –. https://doi.org/10.48550/arXiv.2305.03047
[62] Kuttimani Tamilmani, Nripendra P. Rana, Samuel Fosso Wamba, and Rohita Dwivedi. 2021. The extended Unified Theory of Acceptance and Use of Technology (UTAUT2): A systematic literature review and theory evaluation. *International Journal of Information Management* 57 (4 2021), 102269. https://doi.org/10.1016/j.ijinfomgt.2020.102269
[63] Nicky Terblanche and Martin Kidd. 2022. Adoption Factors and Moderating Effects of Age and Gender That Influence the Intention to Use a Non-Directive Reflective Coaching Chatbot. Issue 2. https://doi.org/10.1177/21582440221096136 Cited by: 33; All Open Access, Gold Open Access.
[64] Nicky Terblanche, Joanna Molyn, Erik de Haan, and Viktor O. Nilsson. 2022. Comparing artificial intelligence and human coaching goal attainment efficacy. , e0270255 pages. Issue 6. https://doi.org/10.1371/journal.pone.0270255
[65] Nicky H D Terblanche, Michelle van Heerden, and Robin Hunt. 2024. The influence of an artificial intelligence chatbot coach assistant on the human coach-client working alliance. *Coaching* 17 (2024), 189 – 206. Issue 2. https://doi.org/10.1080/17521882.2024.2304792 Cited by: 4; All Open Access, Hybrid Gold Open Access.
[66] N H D Terblanche, G P Wallis, and M Kidd. 2023. Talk or Text? the Role of Communication Modalities in the Adoption of a Non-directive, Goal-Attainment Coaching Chatbot. *Interacting with Computers* 35 (2023), 511 – 518. Issue 4. https://doi.org/10.1093/iwc/iwad039 Cited by: 4; All Open Access, Hybrid Gold Open Access.
[67] Tim Theeboom, Bianca Beersma, and Annelies E.M. van Vianen. 2014. Does coaching work? A meta-analysis on the effects of coaching on individual level outcomes in an organizational context. *The Journal of Positive Psychology* 9 (1 2014), 1–18. Issue 1. https://doi.org/10.1080/17439760.2013.837499
[68] Debbe Thompson and Tom Baranowski. 2019. Chatbots as extenders of pediatric obesity intervention: An invited commentary on "feasibility of Pediatric Obesity & Pre-Diabetes Treatment Support through Tess, the AI Behavioral Coaching Chatbot". *Translational Behavioral Medicine* 9 (2019), 448 – 450. Issue 3. https://doi.org/10.1093/tbm/ibz065 Cited by: 13; All Open Access, Hybrid Gold Open Access.
[69] Jules Vandeputte, Antoine Cornuéjols, Nicolas Darcel, Fabien Delaere, and Christine Martin. 2022. Coaching Agent: Making Recommendations for Behavior Change. A Case Study on Improving Eating Habits. , 1292 – 1300 pages. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85134306452&partnerID=40&md5=a144912bc68408fd9a2f15222d8710f1 Cited by: 2.
[70] Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. 2023. Cue-CoT: Chain-of-thought Prompting for Responding to In-depth Dialogue Questions with LLMs.
[71] 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 (2023), 33 – 51. Issue 1. https://doi.org/10.1515/icom-2022-0042 Cited by: 4; All Open Access, Gold Open Access.
[72] Joseph Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM* 9 (1 1966), 36–45. Issue 1. https://doi.org/10.1145/365153.365168
[73] Simon Western. 2017. *The Key Discourses of Coaching*. SAGE reference, Los Angles, 42–61.
[74] George White and Samantha Green. 2024. Heuristic Reasoning in AI: Instrumental Use and Mimetic Absorption. https://doi.org/10.48550/arXiv.2403.09404
[75] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–22. https://doi.org/10.1145/3491102.3517582
[76] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Qin Liu, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huan, and Tao Gui. 2023. The Rise and Potential of Large Language Model Based Agents: A Survey. *ArXiv* abs/2309.07864 (2023), –. https://doi.org/10.48550/arXiv.2309.07864
[77] Hye Sun Yun, Shuo Zhou, Everlyne Kimani, Stefan Olafsson, Teresa K O'Leary, Dhaval Parmar, Jessica Hoffman, Stephen Intille, Michael Paasche-Orlow, and Timothy Bickmore. 2022. Techno-spiritual Engagement: Mechanisms for Improving Uptake of mHealth Apps Designed for Church Members. , 130 – 138 pages. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85128842969&partnerID=40&md5=0cc40eaa6add5fa95c2a6ac6d8856555 Cited by: 0.
[78] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–21. https://doi.org/10.1145/3544548.3581388
[79] Pengyu Zhao, Zijian Jin, and N Cheng. 2023. An In-depth Survey of Large Language Model-based Artificial Intelligence Agents. *ArXiv* abs/2309.14365 (2023), –. https://doi.org/10.48550/arXiv.2309.14365
[80] Asier López Zorrilla and M. Inés Torres. 2022. A Multilingual Neural Coaching Model with Enhanced Long-term Dialogue Structure. *ACM Transactions on Interactive Intelligent Systems* 12 (6 2022), 1–47. Issue 2. https://doi.org/10.1145/3487066

A Online Appendix

This anonymous link will lead to an anonymous OSF page where the online Appendix can be downloaded. Link: https://osf.io/ge236/?view_only=98a4e12431cb4faf976bb4c254b618d5.