

Please quote as: Göldi, A. & Rietsche, R. (2024). Chatbot Agents Displaying Non-factive Reasoning Enhance ExpectationConfirmation. International Conference on Information Systems (ICIS), .

Chatbot Agents Displaying Non-factive Reasoning Enhance Expectation Confirmation

Completed Research Paper

Andreas Göldi

University of St.Gallen
andreas.goeldi@unisg.ch

Roman Rietsche

Bern University of Applied Science
roman.rietsche@bfh.ch

Abstract

Current guidelines suggest setting low expectations for chatbots for continued engagement. That is of little use if only high expectations entice users. Such are raised by recent innovations, chatbot agents which generate final responses not directly via Large Language Models but after multiple intermediate generations of text as reasoning. This differentiates chatbot agents from other chatbots. Since these intermediate steps are generated, what impact has displaying them? This study aims to assess the impact of displaying these intermediate reasoning steps, conceptualizing them as enabling mindreading in the semi-literal sense, namely reading the internal reasoning generated by the agent. In 3 studies (N=280), we examine whether non-factive displays, which present reasoning as belief rather than knowledge, enhance user expectation confirmation. Results show that non-factive reasoning displays significantly improved confirmation. Therefore, displaying non-factive reasoning not only differentiates these agents in competitive markets but also improves user interaction.

Keywords: LLM, chatbot, conversational agent, LLM agent, mindreading

Introduction

The growing use of chatbots across various sectors signifies a notable transformation in human-technology interaction. Initially developed as simple automated responders, chatbots have transformed into essential tools within many areas, including customer support, personal assistance (Lechler et al. 2019; Maedche et al. 2019) and employee onboarding (Ritz et al. 2023). The development of Large Language Models (LLMs) has augmented the functionality of chatbots, such as OpenAI’s ChatGPT or Meta’s LLaMA, enhancing their ability to understand and generate text that closely resembles human communication (Rae et al. 2021).

Recent improvements in chatbot technologies have enabled these systems to facilitate dialogues that closely resemble human interactions (Kusal et al. 2022). A crucial element of human communication involves *mindreading* - the skill to deduce others’ thoughts and intentions. In human-to-human conversation, mindreading is frequently employed, particularly during delayed responses, to interpret the thoughts of counterparts (Westra and Nagel 2021). However, even in human-to-human conversation, mindreading can only ever be guesswork.

Mindreading is not limited to human-to-human conversation, humans attempt to understand how chatbots come up with their responses as well, but so far, non-agent chatbots calling LLMs to generate their answers remained black boxes. These non-agent chatbots, like blenderbot (Roller et al. 2020), generate responses by a single model call. A recent development in AI changes this. Namely, LLM-based AI agents (in this

paper “chatbot agent”), distinguish themselves by their use of a multi-step reasoning process (Zhao et al. 2023). Unlike their predecessors, these agents utilize an iterative method where each output becomes a new input, forming a dynamic chain that refines the final response (Wu et al. 2022). This method simulates an advanced reasoning process. Since each reasoning step is a language generation, this allows insight into the inner workings of the agent even if the LLM itself remains a black box. Reasoning here is understood as the agent’s ability to process inputs through multiple intermediary steps before arriving at a conclusion. This reasoning helps to improve responses, and as the intermediate generations are text, they can be read. In fact, reading these intermediate reasoning steps is akin to mindreading the agent.

Chatbot agents may be necessary to handle challenging domains where conversational expertise is needed. One such domain is coaching. Coaching agents must be capable of very adaptive reasoning, since problems differ between persons. To not rely on the one-pass solution offered by out-of-the-box LLMs, agents that reason in multiple steps can explicitly take in theoretical knowledge to provide more helpful responses. However, systems that take multiple steps to respond take longer and are slightly more expensive. Therefore, such agents must be advertised by highlighting their benefits, which raises high and specific expectations that need to be fulfilled. Such advertising, while necessary, may be counterproductive if the chatbot cannot deliver on the high and specific expectation (Grimes et al. 2021). Expectation Confirmation Theory (ECT) suggests that satisfaction and the continued use of technology are influenced by the alignment of initial user expectations with actual experiences (Bhattacharjee 2001). In the context of chatbot agents, managing and potentially leveraging high expectations can help them make use of their superior response quality compared to non-agent chatbots. Displaying the intermediate steps of reasoning by the chatbot agents provides a transparent window into their functioning, aligning actual performance with user expectations and fostering satisfaction. This transparency should explain the chatbot agents’ operational mechanisms and enhance dialogue coherence, crucial for maintaining user expectations (Gernsbacher and Givón 1995). The visibility of each reasoning step confirms the chatbot’s capabilities and supports expectation confirmation.

To date, research has not yet explored the concept of ‘mindreading’ within the context of LLM chatbot agents. Current studies focus more on LLM technical development than on how they simulate complex cognitive processes (Zhao et al. 2023). There is a notable gap as chatbots become prevalent in high cognitive sectors like education and coaching, where understanding user intents and emotions is crucial (Maedche et al. 2019). While agent-based LLMs with visible multi-step reasoning might enhance interaction quality, the effectiveness and user experience impact of such ‘mindreading’ features are not well studied.

We, therefore, pose the research question: *How does displaying explicit reasoning steps in LLM-based chatbot agents influence user perceptions and expectations?*

Our research methodology involves three experimental phases to address this question. The initial phase establishes and evaluates user expectations concerning AI chat interactions. The subsequent phase sees participants observing videos of chats with chatbot agents displaying reasoning in real-time in two conditions and a control. In the final phase, we test our research model experimentally in a between-participants research design. We apply our ‘mindreading’ concept in a digital coaching context, where the agent assists students in defining and pursuing educational goals, offering continuous motivational support (Sonesh et al. 2015). This context is especially suited for the task as it requires high-quality responses and an assurance that the initial high and specific expectations are maintained while using the chatbot agent.

Our study contributes to the academic field in three significant ways. Firstly, we introduce a new prescriptive concept - displaying real-time reasoning steps to users - in AI conversations. This helps the user since they now do not have to mindread in the figurative sense anymore: they can actually read what the agent reasons about. As a side-effect, this helps to make the longer waiting times introduced by calling the model several times until the final response is displayed less of a burden: it is possible to reduce latency drawbacks in LLM agents by replacing standard loading indicators with live reasoning displays. Mindreading displays may further improve the explanation quality of responses and address the opacity issues inherent in these systems. *Secondly*, we expand on expectation confirmation theory (Bhattacharjee 2001) by examining how coherence (Gernsbacher and Givón 1995) and explanation satisfaction (Hoffman et al. 2018) influence confirmation and are therefore associated with the latter’s benefits for information systems (Bhattacharjee 2001). *Thirdly*, we introduce and validate a new scale for assessing subjective coherence in dialogues

and adapt existing scales for measuring explanation satisfaction, thereby advancing methodologies in setting and evaluating user expectations. We will now explain the domain further, then the main concepts and constructs, before delving into our studies.

Background

This article aims to shed light on how we can make use of recent advances in LLM-based agent architectures. Namely, using the novel fact that to generate their final response, these agents generate the intermediate prompts, we aim to investigate how displaying these steps as “reasoning” of the agent impacts user expectation maintenance. To do so, we will briefly discuss the domain we use to investigate this topic: coaching chatbot agents applying psychologically grounded coaching for better personal study goal pursuit. In this domain, chatbot responses can have severe consequences, as personally important goals are discussed. We then describe how chatbot agents can be used to improve responses and how this enables us to display intermediate reasoning steps. Finally, we will discuss how we aim to embed our investigation into expectation-confirmation theory, and why we consider dialogue coherence and explanation as potential pathways that may allow the display of intermediate reasoning steps to increase expectation confirmation, and thus to enable better advertising of advanced features in the field of chatbot agents.

Application Domain: Coaching Chatbots

Coaching

Coaching, as a developmental practice, is characterized by a personalized relationship where a coach collaborates with a coachee to identify and pursue specific personal goals (Sonesh et al. 2015). This inherently dialogical approach (Armstrong 2012) helps to restructure the coachee’s thoughts, values, and feelings to better align with their personal aspirations (Jonathan Passmore 2018). Because coaching was long sparsely regulated, some practices only recently started to follow theoretical foundations (Seligman 2007). However, if applied with a theoretical foundation, coaching has been shown to bolster personal growth (Sonesh et al. 2015). In educational contexts, coaching has proven effective in increasing metacognition among college students, thereby improving their academic performance and skill development (Howlett et al. 2021). Furthermore, coaching assists students in developing a clear vision of their goals and connecting daily activities with their long-term aspirations (Bettinger and Baker 2014).

Study Goal Pursuit

Study goal pursuit involves a long-term commitment to academic objectives, like obtaining a degree. Successful study goal pursuit is not always guaranteed. Poor academic performance, for instance, tends to decrease adaptive, e.g., mastery-oriented, goal pursuits while increasing avoidance behaviors (Senko and Harackiewicz 2005). In this context, coaching plays a pivotal role by boosting self-concordance and commitment, which are critical for successful goal attainment and subsequently improving well-being (Burke and Linley 2007). Moreover, coaching not only aids in achieving these academic goals but also enhances resilience, and helps mitigate stress and depression (Grant et al. 2009).

Coaching Chatbots

The field of coaching chatbots is rapidly advancing, demonstrating significant potential in scenarios where AI-driven tools assist individuals in achieving their goals. Studies show that these chatbots can be as effective as human coaches (Terblanche et al. 2022). Despite their current limitations in proficiency compared to human coaches, their cost-effectiveness makes them a compelling alternative (Lömkker et al. 2021). Due to the rise in LLMs (Bubeck et al. 2023), there is now discourse on whether coaching chatbots will at some point replace human coaches (Passmore and Tee 2023). Even if not all features of a coaching chatbot need to be generative (Göldi and Rietsche 2023), it is clear that this emerging technology holds promise (Göldi 2024), particularly among younger, digitally native populations, heralding a significant shift in how coaching is conceptualized and delivered (Mai and Rutschmann 2023).

LLM-based AI Agents

LLM-based AI agents, also known as Generative Agents, represent an emerging frontier in artificial intelligence (Xi et al. 2023) that leverages LLMs to sense environments, make decisions, and execute actions (Park et al. 2023). These agents distinguish themselves from traditional AI by their enhanced abilities in natural language processing, knowledge storage, and reasoning, making them highly effective in complex problem-solving scenarios (Zhao et al. 2023). Often, LLMs are used in a single pass, with a prompt to get to the desired text generation (Göldi et al. 2024). LLM-based agents employ a dynamic process that generates and uses such textual outputs as intermediate steps. This involves the AI iteratively prompting itself to produce successive inputs, thus enabling the handling of increasingly complex problems through a reasoned sequence of actions (Göldi and Rietsche 2024). This is analogous to chaining LLM calls to improve quality (Wu et al. 2022), only that the process is autonomous and dynamic. Essentially, these agents use the language model not just to generate direct responses but as a cognitive engine to determine and orchestrate the sequence of steps needed to reach a solution, marking a significant evolution in AI capabilities¹. Since these chatbot agents respond slower than non-agent chatbots, users will only choose agent chatbots if they expect them to perform better.

Maintaining Expectations through the Display of Intermediate Reasoning Steps

Current state-of-the-art advice on setting expectations for chatbots is to keep them low, so that disconfirmation can be avoided (Grimes et al. 2021). This is rather unhelpful in this competitive space, since by keeping expectations low, users do not have any incentive to try out new offerings.

Expectation-Confirmation Theory

Expectation Confirmation Theory (ECT) posits that the alignment between pre-adoption expectations and post-adoption experiences significantly determines user satisfaction and continuance intentions (Bhattacharjee 2001). The theory is supported by substantial empirical evidence showing that satisfaction derived from meeting or exceeding initial expectations strongly influences the decision to continue using a technology (Jeong et al. 2019). In the context of coaching chatbot agents, applying ECT effectively presents unique challenges and opportunities. The current best practice in chatbot deployment suggests managing expectations by setting them low at the outset to avoid the pitfalls of non-confirmation (Grimes et al. 2021). Contrarily, to differentiate their offerings and enhance market competitiveness, providers often promote these agents as advanced, capable of complex interactions, thereby inadvertently raising user expectations. This dichotomy suggests a need for strategies that not only manage but also strategically leverage high expectations to enhance user satisfaction and continuance. In this context, a chatbot's information quality enhances satisfaction and continuance (Ashfaq et al. 2020). Revealing the intermediate reasoning steps of chatbots could serve as a method to maintain expectations, ensuring that high initial expectations are met and confirmed through transparent operations. ECT is well-established, demonstrating the critical role of expectation management in technology and service success (Lin et al. 2009; Wolverson et al. 2020). In the context of AI-driven technologies like coaching chatbots and LLM agents, rather than setting low expectations to prevent disappointment (Grimes et al. 2021), we investigate the more enticing proposition of how to maintain elevated expectations through transparency and showcasing intermediate reasoning processes. This strategy aligns high initial expectations with actual performance, fostering user satisfaction and encouraging continued engagement according to ECT principles. Importantly, we do not test ECT but use its established assumptions to ground our investigation. ECT established that confirmation is essential, the question is now how to establish it for a coaching chatbot agent, pointing the way for all chatbot agents.

Coherence in Dialogue

Chatbots are dialogue machines. Therefore, to maintain expectations on chatbots, it is crucial that the dialogue feels coherent. Coherence in conversation is important for maintaining mental continuity and ensuring the smooth flow of information, which is essential for confirming expectations during interactions

¹A popular agent framework: <https://python.langchain.com/docs/modules/agents/>

(Gernsbacher and Givón 1995). The display of intermediate reasoning steps in dialogues can enhance coherence by making the flow of conversation more predictable and understandable. This predictability allows users to form causal inferences and anticipate explanations, which helps to resolve ambiguities and ensures that the discourse remains cohesive (Rohde et al. 2011). Furthermore, coherence is not just about maintaining a logical sequence but also involves responsiveness and adaptation to the dynamic shifts in dialogue, which includes recognizing and addressing the structure of questions, answers, and shifts in participation (Schiffrin 1985). By enhancing coherence, chatbots can more effectively mirror human conversational characteristics, making interactions not only more engaging but also more satisfying by aligning with users' expectations for a coherent and responsive communication experience (Carfora et al. 2020). This alignment, in turn, supports the confirmation of initial expectations, contributing to a positive user experience and continued engagement. We therefore posit that coherence, as it is subjectively experienced, may partially explain any possible effect of displaying intermediate reasoning steps on confirmation.

Explanation Quality in Dialogue

Explainability is a fundamental attribute for chatbots, fostering understandability, trustworthiness, and a sense of control among users (Laato et al. 2022). These aspects are particularly vital in ensuring that users can effectively interact with and derive meaningful insights from chatbot agents. For coaching chatbots, which are designed to facilitate learning and personal development, the importance of explainability is magnified. Explanation is a transfer of understanding in dialogue (Walton 2004). This underlines the necessity for coaching chatbots to not only respond to queries but to enhance the user's comprehension and mastery of new information. This is supported by findings that higher-quality feedback in dialogues leads to greater educational progress (Aleven et al. 2004), thereby underlining that in the context of coaching, explainability may directly contribute to efficacy and success.

Mindreading in Conversation

Mindreading, the capacity to attribute mental states to others, is a common cognitive ability that underpins everyday conversation, allowing individuals to predict and explain behaviors (Lavelle 2022). This skill, essential for understanding and interacting within social contexts, is scaffolded by linguistic communication, which enhances our ability to interpret and respond to others' mental states efficiently (Fenici and Garofoli 2017). In the context of chatbots, this translates into the capability to process and react to user inputs by attributing intentions and beliefs, thereby facilitating more natural and effective communication (Spaulding 2020). Such basic mindreading processes are crucial even in technology-mediated interactions, mirroring the dynamics of human conversation and improving user experience by aligning responses more closely with user expectations and needs. In conversation, mindreading involves actively monitoring and understanding others' mental states, including what they know or believe, which aids in effective communication. Two main forms of mindreading are distinguished in this context (Westra and Nagel 2021): *factive* and *non-factive* mindreading. *Factive* mindreading focuses on tracking what others know, relying on direct evidence and shared realities without the need to hypothesize about detached beliefs. It is generally easier and more frequently employed in daily conversation. *Non-factive* mindreading, on the other hand, involves attributing beliefs to others that may not align with reality or the mindreader's own beliefs, requiring more cognitive resources and being used less frequently. This distinction is crucial because it affects how conversational interactions are managed, with *factive* mindreading playing a dominant role in regular exchanges by facilitating the mutual sharing of knowledge and minimizing cognitive load. In simple terms, *factive mindreading* refers to "the interlocutor knows" statements, while *non-factive mindreading* refers to "the interlocutor believes" statements. In the context of chatbot agents, the agent will display their reasoning in the I-form, since this display substitutes for the result of mindreading by the user. This shift from speculative inference to directly observable reasoning not only simplifies the user's task but may also reduce cognitive effort, transforming the process from mindreading into straightforward reading.

Building on this background, we formulated the following hypotheses on the variation between no display of intermediate reasoning, the display of *factive*, and of *non-factive* reasoning steps: **(H1)** Confirmation differs between variations; **(H2a)** variations influence coherence and **(H2b)** explanation satisfaction; **(H3a)** coherence and **(H3b)** explanation satisfaction influence confirmation; and **(H4)** confirmation influences

working alliance with the coaching chatbot, as a domain-relevant criterium (Graßmann et al. 2020) for the importance of confirmation.

Artifact

The backend of our coaching chatbot leverages the langchain and langgraph² python libraries to implement an AI-driven reasoning architecture, based on the GPT-4-0125-preview³ model for generating both intermediate steps and the final coaching responses. This setup facilitates a layered interaction process where user inputs undergo multiple stages of processing: initial parsing, context-sensitive reasoning via predefined LLM function calls⁴, and dynamic response generation. At each stage, specific functions categorize user inputs and coaching states. For instance, user messages are analyzed to determine the current coaching stage (e.g., Goal, Reality, Options, Way Forward, Anthony M. Grant 2018) using a custom function call to the model. This analysis is combined with the chat history to generate internal instructions that dictate the next best coaching action. These actions then serve as the final prompts for the GPT model, ensuring that generated final responses are contextually aligned with the user's progression in the coaching session. This multi-tiered process ensures that each interaction is responsive and reflective of an ongoing, evolving coaching dialogue. Importantly, it allows the display of intermediate text generations.

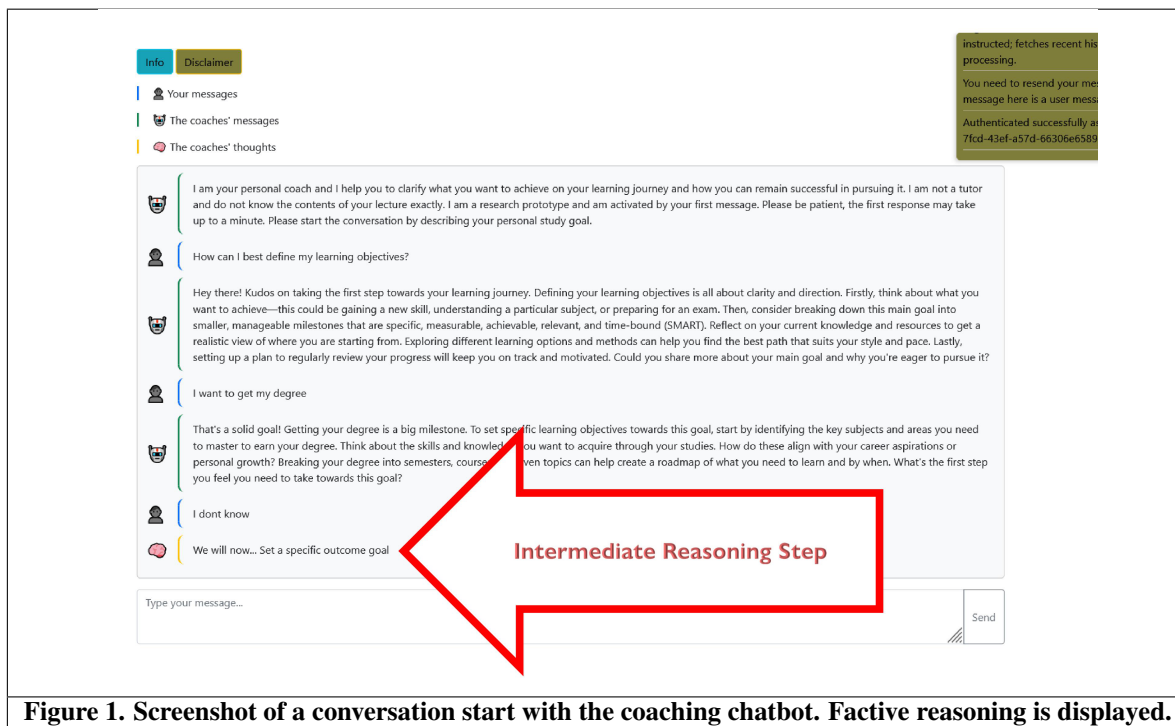


Figure 1. Screenshot of a conversation start with the coaching chatbot. Factive reasoning is displayed.

The front-end of our coaching chatbot is designed as a standard, password-protected chat interface that maintains a continuous WebSocket connection for real-time interaction with users (See Figure 1). Upon successful authentication, users engage in a dialogue where the intermediate reasoning steps are communicated to the user. These reasoning steps are displayed for a minimum of four seconds to ensure the user has adequate time to process the information. The display of these messages is managed by client-side logic, which queues newer messages to replace older ones only after they have been visible for the required duration. Once the chatbot delivers the final response, all reasoning messages are hidden, transitioning the user's focus to the conclusive advice or feedback provided by the chatbot agent. This system enhances user engagement by making the reasoning process transparent and ensuring a fluid and coherent interac-

²<https://python.langchain.com/docs/langgraph/>

³<https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4>

⁴<https://platform.openai.com/docs/guides/function-calling>

tion flow within the coaching session. In this study, we varied the display of intermediate reasoning steps (the coaching chatbot’s “reasoning messages”). In the control group, they remained hidden. In one experimental group, they were prefixed with factive statements about “knowing”, in the other about non-factive statements about merely “believing”.

Reasoning Step	Nonfactive	Factive
Start Thinking	I believe the user stated...	The user stated...
Prediction	I anticipate that the user will...	The user will...
History	I’m trying to recall what was said...	I have to remember what was said...
New BCT	I think this will help...	This will help...
get_GROW_and_BCT_together	I believe the user is...	The user is...
get_GROW_understanding	I’m attempting to understand how the user is...	The user is...
reasoning_path_to_take_now	I think we should now...	We will now...
lookup_of_info	I’m searching for relevant information...	Looking up relevant information...
final_thought	I think this could help...	This will help...

Table 1. Intermediate Reasoning Steps Linked to a Specific Prefix.

Study 1 - Setting and Assessing Expectations

To better understand initial user expectations for chatbots, and specifically for our coaching chatbot agent, we assessed prior experiences and specific expectations. These latter were raised specifically through the textual explanation we provided. To assess these raised expectations, we generated a questionnaire based on the expectations explicitly set in the text.

Automatic Creation of Items to Assess Specific Expectations

To assess specific expectations regarding the coaching chatbot, items were generated using an LLM, specifically GPT-4 Turbo⁵. These items were designed to capture the specific user expectations raised by the textual stimulus, which explained the coaching chatbot and its inner workings (see Table 2 for the generated items).

Method

The exploratory phase included 30 participants (17 female, 12 male, 1 other; $M = 24.6$, $SD = 5.4$), while the replication phase involved another 30 participants (17 female, 13 male; $M = 25.5$, $SD = 4.0$). We compensated the second phase for 5 instead of 3 minutes (paying 9£/hr. on prolific.com) because participants took longer than expected in the first phase. All participants had student status and were fluent in English. Furthermore, because our coaching chatbot is meant to provide help with personal study goals, we required all participants to have one, which was ensured by only including participants that were actively studying towards some degree or certificate.

Stimulus. To set expectations, we wrote a text⁶ explaining the coaching chatbot, with a focus on the multi-step reasoning process that distinguishes it from other similar tools that rely only on the black box of the

⁵System prompt: “Takes in a text and generates a number of questions to be asked in a Likert rating, assessing the explicitly raised expectations with 3 different statements to agree or disagree with each.” The items were then generated verbatim after providing the textual explanation.

⁶The digital coach, accessible after the next page, utilizes advanced coaching practices and artificial intelligence to not only enhance your motivation but also guide you in achieving your study goals. This tool is specially designed to engage you in a dialogue tailored to your needs, whether it’s improving academic performance, acquiring new skills, or deepening knowledge in a specific area.

The core functionality of the coach lies in its sophisticated ‘thinking’ process, which analyzes your messages through multiple stages. Each step is carefully designed to dissect and understand your input, allowing the coach to formulate the most effective, evidence-based responses. This multi-stage reasoning ensures that every piece of advice is not only relevant but also rooted in a deep understanding of your educational challenges and goals.

Designed to support goals set within a 2-3 month timeframe, this tool provides targeted and structured

LLM, and which was the focus of this overall research project. We slightly changed the text in the replication phase to raise expectations more narrowly for a text-based interface. However, these changes did not have any measurable impact besides slightly changing the covariance structure (mean intercorrelation .29 in the exploratory, .36 in the replication phase).

Measures. Besides age and gender, we assessed general self-efficacy (Doll et al. 2021) and prior experiences as well as our main focus of interest, the specific expectations raised by our textual explanation of the coaching chatbot. Experience and expectation items were shown as 7-point Likert scales, with conceptual triples (see Table 2)

Procedure. Participants were invited via prolific.com. After consent, they had to indicate their personal study goal and demographics, as well as their general self-efficacy and experience with professional support, goal setting, and AI chat. Having done so, they were presented with the stimulus text, aimed at setting specific expectations. These expectations were then assessed using the 7-point Likert scales from above.

Category	Item
AI	I believe the use of AI will enhance my motivation.
	AI integration provides significant benefits.
	AI supports my educational needs effectively.
Dialogue	Tailored dialogues improve my academic performance.
	Tailored dialogues help in acquiring new skills.
	Personalized interaction deepens knowledge in specific areas.
Process	Multi-stage reasoning understands my educational challenges.
	Multi-step analysis leads to effective responses.
	Sophisticated thinking provides relevant advice.

Table 2. Specific Expectations (Likert) for Coaching Chatbot Agent.

Results

All scales were consistent, with Cronbach’s alpha between .64 and .93 (average .80). There were no mean differences between the two iterations. Descriptive statistics can be seen in Table 3. To identify key variables influencing expectations, we analyzed feature importance and effect sizes via Random Forest (estimators=100) and Linear Regression models, which revealed that general self-efficacy, expectation of support, and technological expectations were pivotal across the outcomes. Simplified linear regression models were then constructed, focusing on these key predictors for each outcome variable. Upon testing these models on the separate replication dataset, they demonstrated reasonable predictive accuracy, with mean squared errors of 2.60 for AI expectations, 1.93 for dialogue expectations, and 1.65 for process expectations, confirming the relevance of the selected predictors in practical scenarios.

Variable	Mean	SD
General Expectation (Support)	3.79	1.32
General Expectation (Goal)	4.29	1.58
General Expectation (Technology)	4.61	1.20
Expectation of AI	5.03	1.26
Expectation of Dialogue	5.18	1.23
Expectation of Process	5.33	1.11
Self-Efficacy (GSE)	3.85	0.70

Table 3. Descriptive Statistics for Study 1 (pooled data).

guidance, making it an invaluable asset for your educational progress.

Discussion

Study 1 explored the influence of carefully crafted textual explanations on user expectations of a coaching chatbot, utilizing a two-phase approach to validate these influences. The findings underscore the challenges of inducing specific expectations through text, even when leveraging prior user experiences. Despite the structured approach to set and assess expectations - where GPT-4 Turbo helps generate targeted questionnaire items - the study highlights the complex interplay between user backgrounds and the specific information provided.

Study 2 - Maintaining Expectations - Video Mockup

Method

We recruited $n=60$ participants for our second study, 49 of which did not fail attention checks (28 male, 20 female, 1 other; age: $M = 25.40$, $SD = 5.3$). We compensated for 45 minutes (paying 9£/hr. on prolific.com)). Inclusion and exclusion criteria remained as in Study 1.

Procedure. We retained the procedure from Study 1. After this, participants watched a video of the coaching chatbot and filled in the questionnaires described in the earlier section. They did this thrice, with the same questionnaires after each video. Each video corresponds to a variation of the coaching chatbot interface. One video showed an interaction with the control variation (no intermediate displays). The second and third showed intermediate displays of “reasoning messages”, one with factive, the other with non-factive statements. Participants in this study further provided insight into the observed differences between the three variations.

Measures. Besides all measures from Study 1, we assessed participants experience with the coaching chatbot. We assessed *Mental effort* using a Paas Scale (Paas et al. 1994) in random order on how much effort was expended for mindreading the chatbot, understanding the chatbot’s responses, thinking about their own messages, typing, waiting for the next message. Further, we assessed *explanation satisfaction*, with a scale adapted from Hoffman et al. (2018), replacing “service” with “digital coach”. To validate the scale, we also assessed *causability*, how well the explanation allows causal understanding, with the eponymous scale (Holzinger et al. 2020). Further, we assessed *subjective coherence* of the dialogue using our own items (see Table 4), since existing measures we could find were not geared to dialogue with chatbots. The main outcome variable, *confirmation*, was assessed following Bhattacharjee (2001). We also assessed *difference scores* between the initial expectations and retrospective experience scores. Further, we had participants write down how expectations were confirmed and disconfirmed.

The chatbot’s responses were consistently relevant to the conversation topics.
The chatbot accurately maintained the context throughout our conversation.
The transitions between different topics in the conversation were smooth and logical.
The chatbot’s replies built logically upon previous parts of the conversation.
I found the chatbot’s responses to be unpredictable or off-topic at times. (<i>R</i>)
The chatbot effectively referenced earlier parts of the conversation when necessary.
The chatbot’s responses seemed disconnected from the conversation flow. (<i>R</i>)
Table 4. Items for Measuring Subjective Coherence on a 7-point Likert Scale. <i>R</i> = reverse scored. Derivation sources at https://osf.io/s4vk7/?view_only=92f8e8109b6148e09b71e739e9125d9b .

Attention Checks, Understanding Checks, Manipulation Checks. As there have been critiques of using crowd-sourced participants, we built the required checks to ensure data quality. We had different types of checks. Firstly, we used *attention checks*, meaning checks that assessed click-thorough or other undesired, noise-producing behavior. We used both to instruct participants to select specific answer options and retrospectively describe what they had read or seen. Secondly, we used *understanding checks* to assess whether participants understood what the intermediate display meant, namely that it was supposed to indicate the “reasoning” of the coaching chatbot agent. Thirdly, we used *manipulation checks* to indicate whether participants had grasped the essential variation they experienced, namely whether any explanation was provided

for what the coaching chatbot was doing in the period between their message and its response, and whether they had seen a reasoning message and its prefixes (knowing versus believing things about the user). We only excluded participants who failed more than one attention check but accounted for the understanding and manipulation checks in separate analyses.

Results

In the full dataset, descriptive statistics indicated that the mean confirmation ratings across stimuli (control with no display and factive or non-factive displays) were relatively consistent, with means ranging from 3.65 to 3.72 ($SD = 0.88$ to 0.93). Similar patterns were observed for confirmation ratings when coherence and explanation satisfaction were considered as predictors. Filtering for participants that fulfilled all manipulation checks and thus can be assumed to understand the differences in stimuli (to $n=43$, henceforth “completely understood dataset”), confirmation ratings still exhibited comparable means, ranging from 3.53 to 3.70 ($SD = 0.87$ to 0.91). Descriptive statistics for confirmation ratings based on coherence and explanation satisfaction followed similar patterns. Analysis of variance (ANOVA) revealed no significant differences among treatment groups for confirmation ratings in the full dataset ($F(2, 144) = 0.104, p = 0.901$) and the completely understood dataset ($F(2, 126) = 0.468, p = 0.627$).

In the full dataset, a linear regression model showed a significant positive relationship between coherence and confirmation ratings ($\beta = 0.698, p < 0.001$). However, there were no significant interaction effects between stimuli and coherence on confirmation ratings. Further, with explanation satisfaction as the mediator, a significant positive association was observed between explanation satisfaction and confirmation ratings ($\beta = 0.698, p < 0.001$). Again, no significant interaction effects were found between treatment groups and explanation satisfaction on confirmation ratings. In the completely understood dataset (no understanding or manipulation check failed), similar patterns were observed. Coherence significantly predicted confirmation ratings ($\beta = 0.694, p < 0.001$), while no significant interaction effects were detected between treatment groups and coherence on confirmation ratings. Moreover, explanation satisfaction exhibited a significant positive relationship with confirmation ratings ($\beta = 0.725, p < 0.001$), with no significant interaction effects between treatment groups and explanation satisfaction on confirmation ratings.

In exploring the impact of the design variations, a thematic analysis of open-answer feedback uncovered nuanced perceptions among participants regarding the artificial intelligence’s functionality and human-likeness. This analysis was based on responses to the question, “What impact if any had the differences on your perception of the digital coach?” Participants expressed a dichotomy in their experiences: while some found that reasoning messages enhanced the perceived competence of the digital coach (“With thought messages it seemed more competent.”), others felt these elements underscored its mechanical nature, reducing perceived trustworthiness (“The videos with his thoughts made him look dumber, more robotic and less trustworthy than the one without them”). Overall, these findings suggest that while coherence and explanation satisfaction independently influence confirmation ratings, the variation of reasoning displays did not significantly moderate these relationships.

Discussion

The study highlights the robust impact of coherence and explanation satisfaction on user confirmation ratings, underscoring their pivotal role in aligning user experiences with expectations. Despite no significant differences within stimuli, the results affirm that both coherence of dialogue and perceived quality of explanations - whether they frame AI reasoning as “knowing” or “believing” - consistently support confirmation. Within-person effects of the stimulus variations (the different types of reasoning displayed in the videos) however were not strong enough to be detected. This may be due to an absence of a real effect, or due to the nature of video versus real interaction.

Study 3 - Maintaining Expectations - Real Interaction

Method

For this study, we recruited $n=160$ participants (82 male, 77 female; age: $M = 25.94$, $SD = 6.7$). We compensated for 25 minutes (paying 9£/hr. on prolific.com). Inclusion and exclusion criteria remained as in Study 1 and 2.

Procedure. We retained the procedure from Study 1. After this, participants were randomly assigned to one of three groups. They were either in a control group, not receiving any displays of intermediate steps (“reasoning” messages), or they received them in the factive or non-factive condition. Depending on the assignment, participants clicked on different links to the chatbot agent variant and continued the survey after at least 5 minutes.

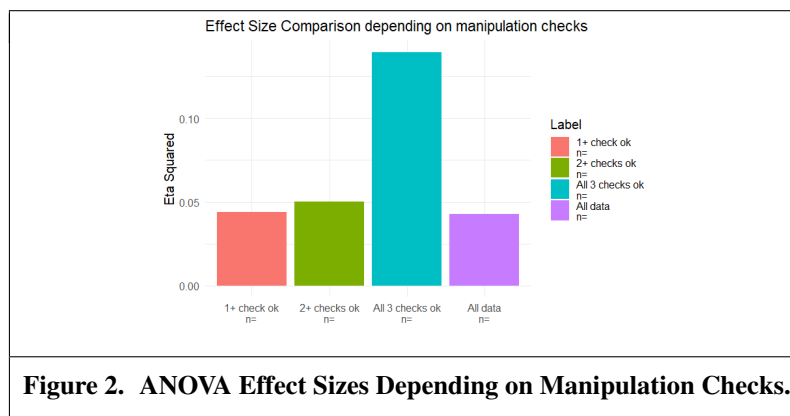
Measures. We retained all measures from Study 2 except for the perception of difference, as participants only came into contact with one variation each. We also assessed the *working alliance* participants felt they had built with the coaching chatbot during their interaction (Munder et al. 2010).

Results

Group	Confirmation	Coherence	ES*	Causability	Self-efficacy	Effort	Working Alliance
Control	3.79 (0.85)	5.41 (0.84)	5.19 (0.83)	4.05 (0.40)	3.92 (0.57)	5.23 (1.33)	3.99 (1.24)
Factive	3.81 (0.78)	5.61 (0.99)	5.27 (1.12)	3.96 (0.61)	4.01 (0.51)	5.48 (1.42)	4.11 (1.15)
Non-factive	4.15 (0.73)	5.65 (0.93)	5.44 (0.81)	4.04 (0.54)	3.89 (0.58)	5.57 (1.57)	4.23 (1.00)

Table 5. Summary of Constructs. *ES = Explanation Satisfaction.

Descriptive statistics can be found in Table 5. An analysis of variance (ANOVA) was conducted to examine the effect of group differences on confirmation. The model was significant, $F(2, 156) = 3.483, p = .033$. Post-hoc comparisons using the Tukey HSD test revealed a non-significant difference between the Non-factive and Factive groups (difference = 0.342, $p = .066$) and between the Control and Non-factive groups (difference = $-0.356, p = .056$). No significant difference was observed between the Control and Factive groups (difference = $-0.014, p = .996$). Repeating the ANOVA with a dataset filtered for all manipulation checks (to $n=87$ participants, again henceforth called “completely understood dataset”) resulted in a significant model as well, $F(2, 84) = 6.817, p = .002$. Post-hoc tests indicated significant differences between the Control and Non-factive groups (difference = $-0.668, p = .002$) and a marginal difference between the Non-factive and Factive groups (difference = 0.545, $p = .048$).



Mediation analyses were performed twice to explore the roles of coherence and explanation satisfaction.

In the first mediation model, coherence did not significantly mediate the relationship between group and confirmation (ACME = 0.0147, 95% CI [-0.132, 0.160], $p = .884$). The direct effect of group on confirmation remained significant (ADE = 0.3276, $p = .004$). The second mediation model involving explanation satisfaction similarly showed no significant mediation (ACME = 0.0891, 95% CI [-0.0809, 0.290], $p = .316$), though the direct effect was significant (ADE = 0.2532, $p = .032$). The repeated mediation analyses with the completely understood dataset showed a non-significant mediation by coherence (ACME = 0.0891, 95% CI [-0.1576, 0.300], $p = .384$) but a significant direct effect (ADE = 0.4559, $p = .012$). Explanation satisfaction also did not significantly mediate the effect (ACME = 0.0778, 95% CI [-0.2263, 0.340], $p = .616$), although the direct effect was significant (ADE = 0.4672, $p < .001$).

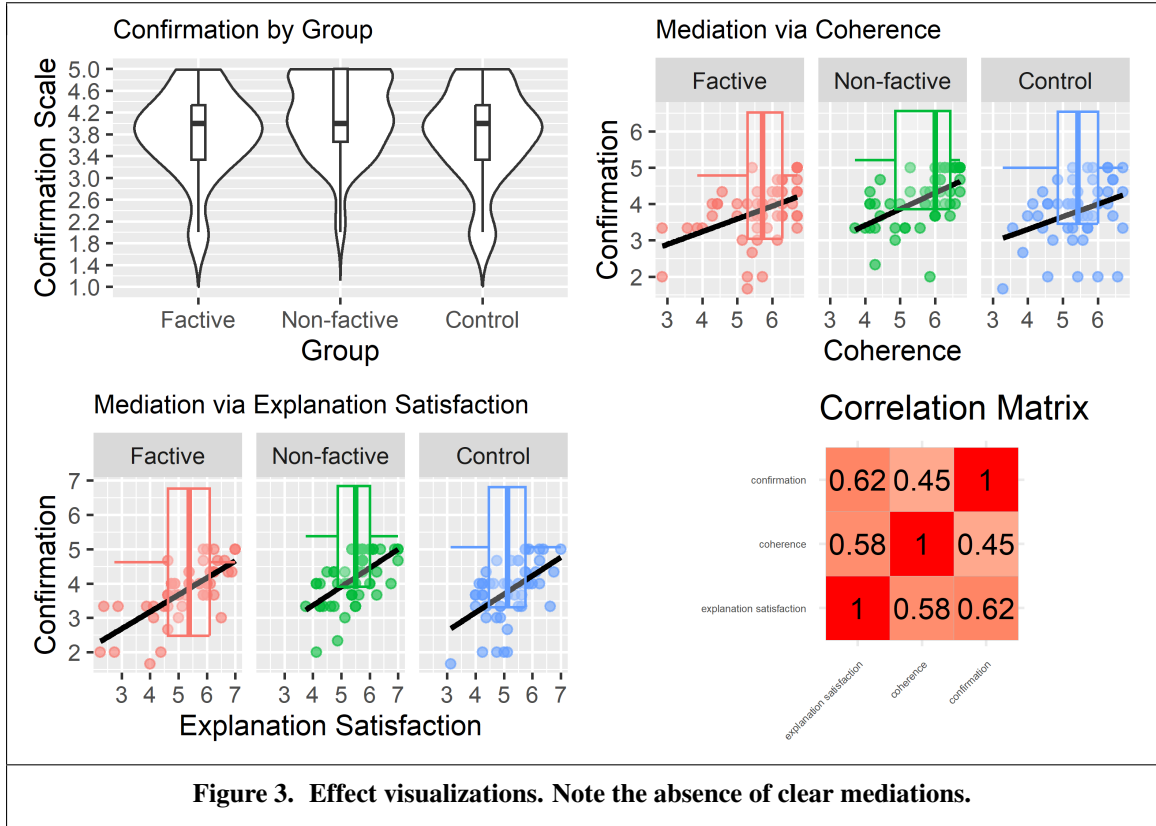
In a full model with all interaction terms incorporating group identity, coherence, and explanation satisfaction, contrasting results emerged between the full dataset and the completely understood dataset. The full dataset displayed no significant interactions, suggesting that noise drowned out the effects. However, the analysis of the completely understood dataset painted a different picture, revealing significant and complex interactions: a positive interaction between group Control and coherence (coefficient = 2.5915, $p = .020$) and between-group Control and explanation satisfaction (coefficient = 2.4878, $p = .032$), alongside a negative interaction involving group Control, coherence, and explanation satisfaction (coefficient = -0.4828, $p = .021$). This subset analysis showed that the interaction terms notably suppressed the main effects of coherence and explanation satisfaction, indicating intricate dynamics specific to this dataset, which substantially contributed to explaining the variance in confirmation outcomes, with an Adjusted $R^2 = 0.4396$.

To validate the importance of confirmation, we also assess working alliance, which is a measure of major importance in the coaching context (Graßmann et al. 2020). Confirmation correlates highly (.5 [0.38, 0.616]) with working alliance (scale measuring the relationship with the coach). The findings suggest that group differences influence confirmation, and these effects are partly mediated by coherence and explanation satisfaction, albeit not consistently across models. The significant interactions in the combined models highlight the complex interplay between group membership and cognitive processes in understanding confirmation. These results underline the importance of considering multiple mediators in exploring the mechanisms underlying group differences in cognitive tasks.

Discussion

This study substantiates the hypothesis that the mode of presenting intermediate reasoning—factive versus non-factive—significantly influences user confirmation, directly linking to the insights from expectation-confirmation theory as outlined in the background. The significant differences between treatment groups, particularly between non-factive and factive presentations in the refined dataset, indicate that it is not just the visibility of AI reasoning that matters but its qualitative character that shapes user expectations. Although coherence and explanation satisfaction did not mediate the relationship between group variations and user confirmation, their significance in the model emphasizes their essential role in aligning user experiences with expectations, as previously suggested in the literature on explainable AI and dialogue coherence. The direct impact of these factors on user satisfaction supports the theoretical underpinnings from sources like Gernsbacher and Givón (1995) and Rohde et al. (2011), who discuss the importance of coherent and understandable interactions in maintaining user satisfaction and engagement. The significant interactions found, particularly involving the control group and factors like coherence and explanation satisfaction, underscore a complex relationship where the context and quality of explanations provided are crucial. This complexity is predictive of the way users process and evaluate AI-generated information, echoing insights from Walton (2004) on the importance of high-quality explanations in educational and developmental contexts.

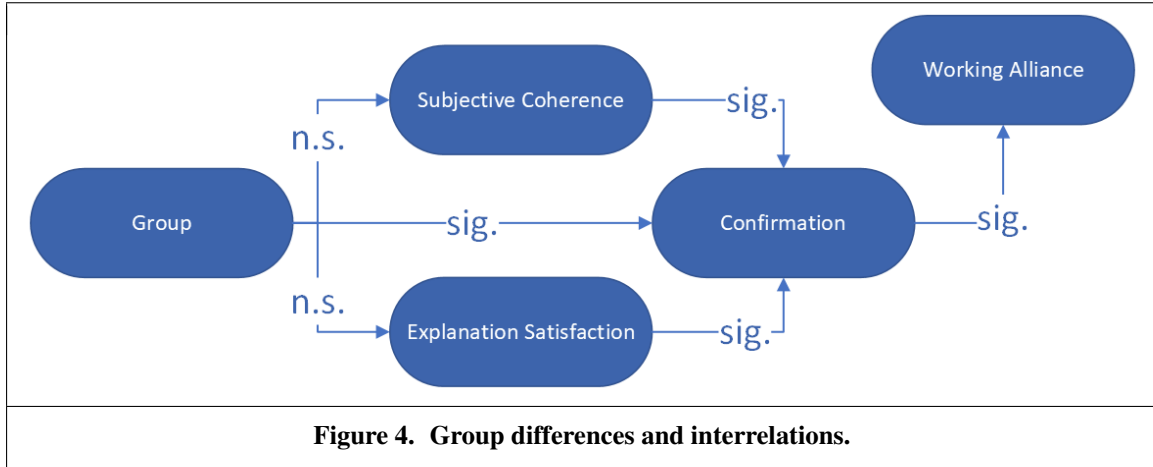
The findings suggest that effective deployment of AI in coaching applications should consider the accuracy of information and how it is cognitively received by users, in line with the theoretical discussions on mindreading in conversation from Lavelle (2022). Future work might expand on these findings by examining individual differences among users that could affect their responsiveness to different types of reasoning presentations. This approach would continue to build on the established theoretical framework and provide deeper insights into optimizing AI interactions for varied user bases.



General Discussion

We investigated the role of intermediate reasoning step displays in LLM-based coaching chatbot agents, focusing on their influence on user expectation satisfaction and coherence, along with the confirmation of those expectations. Unlike other chatbots, chatbot agents generate intermediate reasoning steps in natural language. These steps can be displayed, allowing users to read what in human conversation they would attribute to their interlocutors via mindreading. The key takeaway is that such displays increase expectation confirmation, and that both dialogue coherence as well as expectation satisfaction increase it as well. While we did find significant interactions between groups and expectation satisfaction and coherence in a full model in one dataset, this finding was not consistent throughout studies, with overall indications pointing towards independent direct effects of intermediate reasoning step displays, coherence, and expectation satisfaction on confirmation. This finding can inform both the design of chatbot agents, highlighting that it pays to make use of the intermediate reasoning generated by displaying it. It further enhances our understanding of expectation confirmation by adding the robust effects of coherence and expectation satisfaction to our knowledge of antecedents.

Our results indicate that (*H1*) there are differences in confirmation between no display of intermediate reasoning, the display of factive, and of non-factive reasoning steps. However, this is driven by non-factive reasoning displays, which strengthen confirmation, and only occurs with real interaction, not mere video mock-ups. Neither (*H2a*) coherence, nor (*H2b*) explanation satisfaction mediated this effect significantly, except in our last model when accounting for all manipulation checks. This likely means that if participants are not fully conscious of the nature of the reasoning display, the display has no effect on either coherence or explanation satisfaction. In all cases, however, are (*H3a*) coherence and (*H3b*) explanation satisfaction in-



independently significant to explain confirmation. For the importance of confirmation itself, we can maintain the hypothesis (*H4*) that working alliance, which is crucial for coaching success (Graßmann et al. 2020), is positively related to expectation confirmation after using a coaching chatbot agent. The main findings thus centers on insights on the antecedents of expectation confirmation. Firstly, we can report that for the display of intermediate reasoning steps, it seems to matter how these steps are framed. Namely, non-factive reasoning displays, indicating belief instead of knowledge, were superior to not displaying reasoning steps, while that was generally not the case for factive reasoning displays. The non-factive approach, by not assuming the truth of the chatbot’s statements, potentially reduces user resistance to the AI’s suggestions, thus enhancing the user’s satisfaction and fostering a stronger working alliance.

The role of manipulation checks was crucial in understanding the dataset fully. In the “completely understood dataset”, where participants accurately grasped the variations in stimuli, effect sizes were notably larger (see Figure 4), which also allowed for the suppression of main effects of mediators, as we had expected for the full dataset. This underlines how important it is that users fully grasp the non-factive nature of reasoning-displays. For researchers, we recommend following our manipulation check approach or even extending it. For designers, we recommend obviously non-factive displays of intermediate reasoning, so that assumptions are couched in terms of belief, never knowledge.

The prescriptive knowledge gained from this study informs the design of chatbot agents, suggesting that the display of non-factive reasoning enhances expectation confirmation. This knowledge supports the development of more effective AI-driven coaching tools especially. It also translates to the full realm of LLM-based AI agents that produce intermediate, readable text generations. These agents are increasingly important, be it on their own or in multi-agent teams (Zhao et al. 2023). If agents are used as chatbots, it is important that they display intermediate reasoning in non-factive fashion, so that neither intransparency nor overclaiming, as represented by our control and factive variations, can undermine the effect of displaying them.

We contribute descriptive knowledge on expectation confirmation theory. Firstly, we presented evidence on novel antecedents of confirmation. Namely, coherence and explanation satisfaction. Furthermore, the non-factive display of intermediate reasoning is a further, even more practical antecedent to confirming expectations on chatbot agents.

On the way to our prescriptive and descriptive contributions, we developed new methodology. Namely, we tested a setup where a textual explanation raised specific expectations that can be assessed using LLM-generated items to be rated in a Likert scale. These scales were internally consistent. Secondly, we developed a reliable subjective coherence scale for dialogue with chatbots, which is highly correlated with expectation confirmation.

Limitations and Future research

There are several limitations in our studies. The methodology is constrained by its focused sample, specific use case scenario, specific LLM, and the design of the chatbot agent artifact involved. The results, therefore, may not directly translate to other domains where user expectations and interaction patterns differ significantly. Additionally, the prediction that chatbot agents will become increasingly significant in user interactions is an educated projection rather than a current fact, suggesting a temporal limitation in the assertion of their importance. Future studies could expand the scope of research to include a wider array of underlying LLMs, chatbot functionalities, and user groups, providing a more comprehensive understanding of the dynamics at play. An open question is how the effects of displaying intermediate reasoning relate to those of delays (cf. Harmon et al. 2024). Another avenue for future research is investigating longitudinal effects, especially as concerns chatbot relationship formation and maintenance.

Conclusion

Large Language Models have changed chatbot experiences, with highly competent dialogue and low latencies. A potential next wave of similar proportions, Large Language Model agents, use the model to generate not only final responses but the intermediate steps in readable, generated text. This allows it to display the intermediate steps. This paper shows that in chatbot applications, this display can be shaped to enhance expectation confirmation by framing the intermediate reasoning steps as non-factive, that is, belief-indicating reasoning, which users can read as if mindreading the agents.

References

- Aleven, V., Ogan, A., Popescu, O., Torrey, C., and Koedinger, K. 2004. "Evaluating the Effectiveness of a Tutorial Dialogue System for Self-Explanation," in *Intelligent Tutoring Systems*, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, J. C. Lester, R. M. Vicari, and F. Paraguaçu (eds.). Vol. 3220. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 443–454.
- Anthony M. Grant 2018. "Goals and coaching: an integrated evidence-based model of goal-focused coaching and coaching psychology," in *Handbook of coaching psychology*, S. Palmer and A. Whybrow (eds.). London: Routledge, pp. 34–50.
- Armstrong, H. 2012. "Coaching as dialogue: Creating spaces for (mis) understandings," *International Journal of Evidence Based Coaching and Mentoring* (10:1), pp. 33–47.
- Ashfaq, M., Yun, J., Yu, S., and Loureiro, S. M. C. 2020. "I, Chatbot: Modeling the determinants of users' satisfaction and continuance intention of AI-powered service agents," *Telematics and Informatics* (54), p. 101473.
- Bettinger, E. P. and Baker, R. B. 2014. "The Effects of Student Coaching," *Educational Evaluation and Policy Analysis* (36:1), pp. 3–19.
- Bhattacharjee, A. 2001. "Understanding Information Systems Continuance: An Expectation-Confirmation Model," *MIS Quarterly* (25:3), p. 351.
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. 2023. *Sparks of Artificial General Intelligence: Early experiments with GPT-4*.
- Burke, D. and Linley, P. A. 2007. "Enhancing goal self-concordance through coaching," *International Coaching Psychology Review* (2:1), pp. 62–69.
- Carfora, V., Di Massimo, F., Rastelli, R., Catellani, P., and Piastra, M. 2020. "Dialogue management in conversational agents through psychology of persuasion and machine learning," *Multimedia Tools and Applications* (79:47-48), pp. 35949–35971.
- Doll, E. S., Nießen, D., Schmidt, I., Rammstedt, B., and Lechner, C. M. 2021. *General Self-Efficacy Short Scale-3 (GSE-3)*.
- Fenici, M. and Garofoli, D. 2017. "The biocultural emergence of mindreading: integrating cognitive archaeology and human development," *Journal of Cultural Cognitive Science* (1:2), pp. 89–117.

- Gernsbacher, M. A. and Givón, T. 1995. *Coherence in Spontaneous Text*, vol. 31. Amsterdam: John Benjamins Publishing Company.
- Göldi, A. 2024. "Flexibility in Chatbot Identity Perception," in *AMCIS 2024 Proceedings*, 2024.
- Göldi, A. and Rietsche, R. 2023. "Where to for Automated Coaching Conversation: Structured Intervention or Adaptive Generation?," in *ECIS 2023 Research-in-Progress Papers*, 2023.
- Göldi, A. and Rietsche, R. 2024. "Insert-expansions for Large Language Model Agents," in *AMCIS 2024 Proceedings*, 2024.
- Göldi, A., Wambsganss, T., Neshaei, S. P., and Rietsche, R. 2024. "Intelligent Support Engages Writers Through Relevant Cognitive Processes," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, ACM, 2024, pp. 1–12.
- Grant, A. M., Curtayne, L., and Burton, G. 2009. "Executive coaching enhances goal attainment, resilience and workplace well-being: a randomised controlled study," *The Journal of Positive Psychology* (4:5), pp. 396–407.
- Graßmann, C., Schölmerich, F., and Schermuly, C. C. 2020. "The relationship between working alliance and client outcomes in coaching: A meta-analysis," *Human Relations* (73:1), pp. 35–58.
- Grimes, G. M., Schuetzler, R. M., and Giboney, J. S. 2021. "Mental models and expectation violations in conversational AI interactions," *Decision Support Systems* (144), p. 113515.
- Harmon, K., Lee, H., Khasraghi, B., Parmar, H., and Walden, E. 2024. "Delays in Information Presentation Lead to Brain State Switching, Which Degrades User Performance, and There May Not Be Much We Can Do about It," *MIS Quarterly* (48) 1 2024, pp. 273–298.
- Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. 2018. *Metrics for Explainable AI: Challenges and Prospects*.
- Holzinger, A., Carrington, A., and Müller, H. 2020. "Measuring the Quality of Explanations: The System Causability Scale (SCS): Comparing Human and Machine Explanations," *Kunstliche intelligenz* (34:2), pp. 193–198.
- Howlett, M. A., McWilliams, M. A., Rademacher, K., O'Neill, J. C., Maitland, T. L., Abels, K., Demetriou, C., and Panter, A. T. 2021. "Investigating the Effects of Academic Coaching on College Students' Metacognition," *Innovative Higher Education* (46:2), pp. 189–204.
- Jeong, J. Y., Park, J., and Hyun, H. 2019. "The Role of Emotional Service Expectation Toward Perceived Quality and Satisfaction: Moderating Effects of Deep Acting and Surface Acting," *Frontiers in psychology* (10), p. 321.
- Jonathan Passmore 2018. "Behavioural Coaching," in *Handbook of coaching psychology*, S. Palmer and A. Whybrow (eds.). London: Routledge, pp. 99–107.
- Kusal, S., Patil, S., Choudrie, J., Kotecha, K., Mishra, S., and Abraham, A. 2022. "AI-Based Conversational Agents: A Scoping Review From Technologies to Future Directions," *IEEE Access* (10), pp. 92337–92356.
- Laato, S., Tiainen, M., Najmul Islam, A., and Mäntymäki, M. 2022. "How to explain AI systems to end users: a systematic literature review and research agenda," *Internet Research* (32:7), pp. 1–31.
- Lavelle, J. S. 2022. *Mindreading and Social Cognition*, Cambridge University Press.
- Lechler, R., Stöckli, E., Rietsche, R., and Uebersnickel, F. 2019. "Looking Beneath the Tip of the Iceberg: The Two-Sided Nature of Chatbots and their Roles for Digital Feedback Exchange," in *Proceedings of the European Conference on Information Systems (ECIS)*, Stockholm, Sweden, 2019.
- Lin, C.-P., Tsai, Y. H., and Chiu, C.-K. 2009. "Modeling Customer Loyalty from an Integrative Perspective of Self-Determination Theory and Expectation–Confirmation Theory," *Journal of Business and Psychology* (24:3), pp. 315–326.
- Lömker, M., Weber, U., and Moskaliuk, J. 2021. *Chatbots im Coaching: Chancen im lösungs-fokussierten Coaching*, 1st ed. essentials. Wiesbaden: Springer Fachmedien.
- Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S., and Söllner, M. 2019. "AI-Based Digital Assistants," *Business & Information Systems Engineering* (61:4), pp. 535–544.
- Mai, V. and Rutschmann, R. 2023. "Chatbots im Coaching. Potenziale und Einsatzmöglichkeiten von digitalen Coaching-Begleitern und Assistenten," *Organisationsberatung, Supervision, Coaching* (30:1), pp. 45–57.
- Munder, T., Wilmers, F., Leonhart, R., Linster, H. W., and Barth, J. 2010. "Working Alliance Inventory-Short Revised (WAI-SR): psychometric properties in outpatients and inpatients," *Clinical psychology & psychotherapy* (17:3), pp. 231–239.

- Paas, F. G., van Merriënboer, J. J., and Adam, J. J. 1994. "Measurement of cognitive load in instructional research," *Perceptual and motor skills* (79:1 Pt 2), pp. 419–430.
- Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. 2023. *Generative Agents: Interactive Simulacra of Human Behavior*.
- Passmore, J. and Tee, D. 2023. "Can Chatbots replace human coaches? Issues and dilemmas for the coaching profession, coaching clients and for organisations," *The Coaching Psychologist* (19:1), pp. 47–54.
- Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., Rutherford, E., Hennigan, T., Menick, J., Cassirer, A., Powell, R., van Driessche, G. den, Hendricks, L. A., Rauh, M., Huang, P.-S., Glaese, A., Welbl, J., Dathathri, S., Huang, S., Uesato, J., Mellor, J., Higgins, I., Creswell, A., McAleese, N., Wu, A., Elsen, E., Jayakumar, S., Buchatskaya, E., Budden, D., Sutherland, E., Simonyan, K., Paganini, M., Sifre, L., Martens, L., Li, X. L., Kuncoro, A., Nematzadeh, A., Gribovskaya, E., Donato, D., Lazaridou, A., Mensch, A., Lespiau, J.-B., Tsimpoukelli, M., Grigorev, N., Fritz, D., Sottiaux, T., Pajarskas, M., Pohlen, T., Gong, Z., Toyama, D., d'Autume, C. d. M., Li, Y., Terzi, T., Mikulik, V., Babuschkin, I., Clark, A., Casas, D. d. L., Guy, A., Jones, C., Bradbury, J., Johnson, M., Hechtman, B., Weidinger, L., Gabriel, I., Isaac, W., Lockhart, E., Osindero, S., Rimell, L., Dyer, C., Vinyals, O., Ayoub, K., Stanway, J., Bennett, L., Hassabis, D., Kavukcuoglu, K., and Irving, G. 2021. *Scaling Language Models: Methods, Analysis & Insights from Training Gopher*.
- Ritz, E., Donisi, F., Elshan, E., and Rietsche, R. 2023. "Artificial Socialization? How Artificial Intelligence Applications Can Shape A New Era of Employee Onboarding Practices," in *Proceedings of the 56th Hawaii International Conference on System Sciences*.
- Rohde, H., Levy, R., and Kehler, A. 2011. "Anticipating explanations in relative clause processing," *Cognition* (118:3), pp. 339–358.
- Roller, S., Dinan, E., Goyal, N., Da Ju, Williamson, M., Liu, Y., Xu, J., Ott, M., Shuster, K., Smith, E. M., Boureau, Y.-L., and Weston, J. 2020. *Recipes for building an open-domain chatbot*.
- Schiffrin, D. 1985. "Conversational Coherence: The Role of Well," *Language* (61:3), p. 640.
- Seligman, M. E. P. 2007. "Coaching and positive psychology," *Australian Psychologist* (42:4), pp. 266–267.
- Senko, C. and Harackiewicz, J. M. 2005. "Regulation of Achievement Goals: The Role of Competence Feedback," *Journal of Educational Psychology* (97:3), pp. 320–336.
- Sonesh, S. C., Coultas, C. W., Lacerenza, C. N., Marlow, S. L., Benishek, L. E., and Salas, E. 2015. "The power of coaching: a meta-analytic investigation," *Coaching: An International Journal of Theory, Research and Practice* (8:2), pp. 73–95.
- Spaulding, S. 2020. "What is mindreading?," *Wiley interdisciplinary reviews. Cognitive science* (11:3), e1523.
- Terblanche, N., Molyn, J., Haan, E. de, and Nilsson, V. O. 2022. "Comparing artificial intelligence and human coaching goal attainment efficacy," *PloS one* (17:6), e0270255.
- Walton, D. 2004. "A new dialectical theory of explanation," *Philosophical Explorations* (7:1), pp. 71–89.
- Westra, E. and Nagel, J. 2021. "Mindreading in conversation," *Cognition* (210), p. 104618.
- Wolverton, C. C., Hirschheim, R., Black, W. C., and Burleson, J. 2020. "Outsourcing success in the eye of the beholder: Examining the impact of expectation confirmation theory on IT outsourcing," *Information & Management* (57:6), p. 103236.
- Wu, T., Terry, M., and Cai, C. J. 2022. "AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts," in *CHI Conference on Human Factors in Computing Systems*, S. Barbosa, C. Lampe, C. Appert, D. A. Shamma, S. Drucker, J. Williamson, and K. Yatani (eds.). New York, NY, USA: ACM, pp. 1–22.
- Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., Yin, Z., Dou, S., Weng, R., Cheng, W., Zhang, Q., Qin, W., Zheng, Y., Qiu, X., Huang, X., and Gui, T. 2023. *The Rise and Potential of Large Language Model Based Agents: A Survey*.
- Zhao, P., Jin, Z., and Cheng, N. 2023. *An In-depth Survey of Large Language Model-based Artificial Intelligence Agents*.