Substance over Style: Evaluating Proactive Conversational Coaching Agents

Vidya Srinivas[1]*, Xuhai Xu[2], Xin Liu[1,2], Kumar Ayush[2], Isaac Galatzer-Levy[2], Shwetak Patel[1,2], Daniel McDuff[1,2], Tim Althoff [1,2],
[1]University of Washington, [2]Google Research
vysri@cs.washington.edu, {orson, xliucs, dmcduff, althoff}@google.com

Abstract

While NLP research has made strides in conversational tasks, many approaches focus on single-turn responses with well-defined objectives or evaluation criteria. In contrast, coaching presents unique challenges with initially undefined goals that evolve through multi-turn interactions, subjective evaluation criteria, and mixed-initiative dialogue. In this work, we describe and implement five multi-turn coaching agents that exhibit distinct conversational styles, and evaluate them through a user study, collecting first-person feedback on 155 conversations. We find that users highly value core functionality, and that stylistic components in absence of core components are viewed negatively. By comparing user feedback with third-person evaluations from health experts and an LM, we reveal significant misalignment across evaluation approaches. Our findings provide insights into design and evaluation of conversational coaching agents and contribute toward improving human-centered NLP applications.

Figure 1: **Example Coaching Conversation.** Different coaching behaviors are present throughout the conversation (highlighted in red) based on user responses, corresponding to recommended coaching behaviors (see Section 2). Full transcripts in Section A.7

1 Introduction

Recent NLP research has focused on evaluating tasks in conversational settings, including proactive information-seeking (Li et al., 2024), multi-turn conversational dialogue (Yi et al., 2024; Kwan et al., 2024; Bai et al., 2024), and others. While significant progress has been made, the focus has largely been on exercises that represent one or multiple of the following properties: a single, unchanging, clearly communicated goal, a single correct answer, a high-quality answer expressed in a single turn or utterance, an assumption that objective quality can be evaluated by third parties (e.g., experts or language models), or a clear interaction structure (e.g., all user or model-initiated). However, proactive human-centered NLP applications such as personalized target-guided dialogues (Deng et al., 2024) often exhibit several, concurrent, contrastive challenges. This work focuses on coaching conversations, which differ from the aforementioned tasks with the presence of: 1) open-ended multi-turn interactions, 2) initially ambiguous task definition that must be informed through multiple user interactions, 3) multiple and changing goals that must be refined and prioritized, 4) potentially off-topic user input throughout long-form conversation, 5) varied conversational styles and preferences, 6) mixed-initiative interactions, in which a coach must balance addressing user goals, needs and empowerment, 7) unspoken user needs and incomplete context, and the 8) absence of a single, objectively correct answer (instead relying primarily on subjective first-person evaluation).

As a result, coaching remains difficult and underexplored for proactive agents. In this work, we aim to 1) determine essential components of LM agent

*Work done during an internship at Google

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 20848–20880
July 27 - August 1, 2025

coaching, 2) describe and evaluate LM coaching agents that express distinct coaching styles (from common directive and interrogative to facilitative styles informed by human coaching experts), and 3) better understand agreement or discrepancies between first and third person evaluations of such conversations to inform future approaches. We use health coaching as an illustrative example to frame the coaching domain. Through our study, we find that questioning-only conversation styles typically reduce user engagement and satisfaction. We also find that users place high value on core functional components, which we refer to as substance, and perceive stylistic elements negatively when this substance is lacking. Finally, we find that LMs perform poorly as auto-raters for human-centered subjective tasks, but demonstrate reasonable agreement with users for objective human-centered tasks. Our contributions are as follows:

- We synthesize the expertise and recommendations of health coaching experts ($N = 11$) across a variety of coaching domains to define key capabilities for open-ended coaching.
- We identify LM coaching characteristics and create five conversational multi-turn coaching agents that incorporate different levels of these characteristics to embody directive, interrogative, and facilitative coaching styles
- We conduct a user study ($N = 31$, 50 person hours of feedback, 155 conversations) where participants engage in unsupervised discussions on self-directed, open-ended coaching topics with these agents. We collect first-person feedback.
- We examine the alignment between first-person feedback and third-person annotations from health experts and an LM, highlighting shortcomings of evaluation approaches in this setting.

2 Initial Health Expert Interviews

Health coaches provide general guidance to a user who is looking to pursue a health goal. To inform the design of our coaching agents across general coaching domains, we conducted a formative study with 11 coaches across domains of sleep, fitness, and life coaching. These coaches, or experts, possessed advanced degrees in their respective fields, coupled with substantial professional experience ranging from 4 to 46 years. The one-hour interview consisted of two parts—open-ended questions through which experts shared their expertise on coaching best-practices, and a case study, in which experts reviewed, turn-by-turn, a conversation between a real end-user and a single-prompt LM.

2.1 Health Expert Insights

Based on our interviews, we identified **conversational expert insights** which we classified into two categories: *style* and *substance*. *Style* refers to how something is presented—-its form or delivery. This includes language, tone, and overall impression. *Substance* refers to core components of a helpful and actionable coaching experience, including exploration, context gathering, goal-seeking, and relevant recommendation.

- `I1` **Goal and purpose understanding** (*substance*): Understanding goals and motivations is important. Coaches maintain a goal-oriented conversation and bring clients back to the main goal when they deviate off-topic.
- `I2` **Context clarification** (*substance*): Coaches collect and understand clients' constraints, preferences, and prior resolution attempts to better personalize recommendations.
- `I3` **Relevant recommendations** (*substance*): Coaches provide recommendations that are relevant, actionable, and context-sensitive.
- `I4` **Feedback seeking** (*substance*): Coaches solicit user feedback and update recommendations as necessary based on this feedback.
- `I5` **Active listening** (*style*): Coaches provide occasional reflections to ensure correct understanding and goal alignment promoting client engagement and trust.
- `I6` **Client empowerment** (*style*): Coaches build trust with their clients and guide them to navigate their own paths when exploring potential solutions. Coaches are supportive when clients are moving in a good direction.

2.2 Turn-by-Turn Expert Evaluation

The single prompt used to generate example conversations for the turn-by-turn part of the health expert interviews contained both control flow decisions (i.e. when to change topic, when to ask questions, etc.) and coaching instructions. From expert annotations, we identified cases of premature recommendation and fixated questioning, in which the LM either made recommendations before seeking relevant information or continued down a single, narrow line of questioning without considering additional factors (expert quotes available in Section A.1). These health expert interviews revealed consistent, significant shortcomings of static

LM prompting approaches. These issues, along with the expert insights, informed our designs of conversational flow (Section 4.2).

3 Conversational Paradigms

Before developing a coaching agent that embodies health expert insights, we recognize two conversational paradigms that already exist between humans and LMs, and point out how *expert insights emphasize a third, distinct form of conversation*.

Interrogative interaction occurs when either a human or an LM provides one-sided questioning in a conversation while the opposite party only provides answers; a recommendation is potentially made at the end of the conversation by the questioning party. Such an approach maximizes knowledge gain but minimizes interaction. Questionnaires have always been used for general information seeking, but with the introduction of LMs to interpret and understand responses, proactive QA for open-ended information seeking often embodies this paradigm (Hu et al., 2024; Zhang et al., 2024; Li et al., 2023).

Directive interaction is a paradigm in which the LM takes an active role in providing constant solutions, instructions, and feedback to help a user achieve a goal. The most common examples are today's instruction-tuned LMs, such as ChatGPT, Claude, and Gemini (Team et al., 2023), among others. These models strive to provide continuous guidance and recommendations to a user. This is also a paradigm in coaching (Gabriele et al., 2011; Hammond and Moore, 2018).

Facilitative interaction is a paradigm in which an LM focuses on guiding individuals towards their own solutions rather than providing immediate answers or recommendations. This interaction mode is emphasized in coaching (Schwarz and Davidson, 2008; Schwarz et al., 2011), but is not inherent to today's LMs. Through our formative study Section 2, we identify a central tension in open-ended conversational scenarios, specifically between making a recommendation and gathering further context from the user (Ikemoto et al., 2018).

4 Design Motivation and Components

We introduce intended LM characteristics that can be varied to create different coaches. These are not to be confused with *expressed* or *perceived* characteristics. *Expressed* characteristics are those that are evaluated by a third party. We show alignment of expressed and intended characteristics in Section 7.6. *Perceived* characteristics, are evaluated, post-conversation, in Section 6. This differentiation can be critical for conversational constructs, including empathy (Sharma et al., 2020, 2022). Two modules encode different levels of coaching expertise and three agentic conversational flow modules allow us to implement the paradigms in Section 3. All agents use Gemini 1.5 Pro as a base LM.

4.1 Coaching Expertise Variations

We introduce two coaching modules, `Base`, and `Expert`, which vary *style* components as defined in Section 2.1. The `Base` and `Expert` modules share core instructions for proactive questioning; however, they differ in specific conversation objectives and conversation flow instructions. One can think of the `Base` module as a coach with good intention but without formal training. In particular, the `Base` module specifically outlines proactive questioning guidelines, but does not define specific topics, such as constraints or preferences, to talk about or heavily emphasize motivational language (see Section A.4 for all prompts).

In contrast, one can think of the `Expert` module as an experienced coach. The `Expert` module explicitly outlines a goal-oriented questioning line, first breaching goal and purpose, and then constraints, preferences, and barriers before making a recommendation. The `Expert` module (see Section A.4 for prompt) emphasizes motivational behavior, explicitly incorporating active listening (`I5`) and client empowerment (`I6`).

4.2 Conversation Flow Variations

We introduce *interrogative*, *directive*, and *facilitative* conversation flows, which vary in *substance* components as defined in Section 2.1. We first observe that the *directive* flow is trivial to implement through a simple, single prompt. Our base LM, Gemini, inherently embodies a recommendation-heavy, suggestion-first flow (`I3`) as it is instruction-tuned to do so. It does not place any explicit emphasis on exploratory behavior (`I1`), context gathering (`I2`), or feedback seeking (`I4`).

While some coaching styles are easily elicited through simple prompts, others are harder to implement because it is difficult to specify through a single prompt when to make control flow decisions. These decisions include factors such as when to stop questioning, when to make a recommendation, or when to end a conversation. Deciding when to make such decisions often requires further, parallel control decisions to be considered. To make these decisions, we introduce explicit conversation flow, which defines a chain of thought (Wei et al., 2022) before each agent response.

Explicit conversation flow modules steer conversation direction, executing control decisions in parallel during each turn. Each module consists of two LM inferences using Gemini 1.5 Pro. The first inference runs after each user utterance and outputs a binary decision. The second inference specifies how to perform the LM response to the user and executes based on this binary decision. This allows all first inferences to be run in parallel, while only one second inference will execute. When multiple first inferences return positive decisions, the first module in the conversation flow is prioritized (questioning before recommendation etc.). This allows the furthering of conversation to be prioritized over conclusion of the conversation.

The `Probing` module is introduced to ask further clarification questions when something is unclear, vague, or when the user seems unsure or confused. The `Recommendation` module is introduced to determine when to make a recommendation and to seek user feedback. The `Resolution` module is introduced to determine when a conversation has reached a reasonable conclusion. All prompts are presented in Section A.4. We use a combination of the `Probing` and `Recommendation` modules to implement the `Interrogative` conversational flow (emphasizing context gathering (`I2`), and de-emphasizing goal and purpose understanding (`I1`), feedback seeking (`I4`), and suggestion frequency (`I3`)) and a combination of the `Recommendation` and `Resolution` agents to implement the `Facilitative` conversational flow, emphasizing goal and purpose understanding (`I1`), context gathering (`I2`), and feedback seeking (`I4`), and exhibiting medium suggestion frequency (`I3`)). We discuss how these styles can be combined to produce LM coaches in Section 5.

5 Coaching Agent Design

Using these components we design five agents. The `Base-Interrogative` and `Expert-Interrogative` agents are implemented with one coaching module, `Base` or `Expert` (Section 4.1), and `Interrogative` conversational flow. The coaching module controls foundational proactive coaching while the `Interrogative` conversational flow emphasizes knowledge gain before a recommendation is made.

The `Base-Facilitative` and `Expert-Facilitative` agents are implemented with one coaching module, `Base` or `Expert`, and `Facilitative` conversational flow. The coaching module controls foundational proactive coaching, while the `Facilitative` flow enforces timely recommendation and graceful conclusion. We implement the directive coaching style in the `Directive` agent with a prompt that asks the LM to act as a conversational health coach. We do not additionally prompt this agent or add control flow as it already embodies the coaching paradigm that we mean to test in Section 4.1 and Section 3.

5.1 Hypotheses for Agent Behavior

We make the following hypotheses about agent behavior based on intended agent behaviors and level of alignment with expert insights:

- `H1` (**Style**): Agents with an *expert* coaching module will perform better than corresponding agents with a *base* coaching module for active listening (`I5`) and client empowerment (`I6`), with the `Directive` agent performing in between.
- `H2` (**Substance**): `Facilitative` agents will perform better than `Directive` agents, which will perform better than `Interrogative` agents on the following: goal purpose and understanding (`I1`), context clarification (`I2`), relevant recommendations (`I3`), and feedback seeking (`I4`).
- `H3` (**Overall**): Overall, in terms of alignment with expert insights across all indicators (`I1`–`I6`), agent performance is expected to decrease in the following order (from best to worst): `Expert-Facilitative`, `Base-Facilitative`, `Directive`, `Expert-Interrogative`, `Base-Interrogative`.

In Section 7, we systematically evaluate these agents. In Section 7.6, we validate these hypotheses through expert evaluation.

6 Research Questions & Study Design

In this section, we describe a user study evaluation of our five proposed coaching agents, to answer the following research questions:

RQ1: Which agents do users prefer overall and do they have specific characteristics that stand out?
RQ2: What matters more in a coaching conversational agent, *substance* or *style*?
RQ3: Are LMs competent coaches?
RQ4: What are the factors that drive user satisfac-

tion across agents?
RQ5: Can LM auto-raters capture first-person user preferences and third-person coach expertise?
RQ6: Do agent LM coaches demonstrate intended behavior?

6.1 End-User Evaluation Study

We recruited 37 English-speaking participants, 31 of whom participated in our study. Participants varied in age (18 in range 25-34, 12 in range 35-44, and 1 in range 55-64) and self-identified gender (19.3 % female, 81.7% male). Each participant attended a 1.5-hour session, during which they interacted with and evaluated the five agents in an order denoted by a balanced Latin square. To guide interactions, we curated 33 open-ended scenarios reflecting common concerns across sleep, fitness, repeated habits, and everyday situations (Table 6). Participants selected the scenario that best aligned with personal interests to ensure a realistic coaching use case. Participants were instructed to treat each agent interaction independently, providing consistent background information as if consulting a new expert. Each interaction began with the same initial sentence from the chosen scenario, after which participants conversed freely. Conversations concluded naturally—-either with a resolution or recommendation, or when further discussion was deemed unproductive. After each interaction, participants assessed the agent individually. At the end of the study, the users ranked all five agents on different metrics, discussed further in Section 7.

A meaningful evaluation requires users to be invested in such a study beyond participation compensation. A key strength of our study is the flexibility it offers participants in scenario selection, ensuring relevance to personal behavioral health goals. Additionally, by allowing open-ended interactions without rigid guidelines on conversation structure, our study captures a range of realistic user experiences and human-language model interactions that may be overlooked in constrained studies or automated evaluations.

7 User Study Results & Discussion

In this section, we describe the results of our user study. We primarily collected three data types: Likert scale ratings, agent rankings, and open-ended feedback. All Likert scale ratings are on a scale of 1-5 (1–worst, 5–best). When applicable, quotes are specified along with participant number.

7.1 Which agents do users prefer overall?

Method. We define win rate as the combined percentages of first and second rankings of agents. We also collected participant ratings on a 5-point scale for self-identified LM chatbot usage experience and experience working with human health coaches. A rating of 1 indicated no experience; a rating of 5 indicated extensive experience.

Results. We observe that 67.7% of users rank facilitative agents as their top preference, followed by directive, and then interrogative agents (Figure 2). The `Expert-Facilitative` accounts for 41.9% of first rankings and `Base-Facilitative` accounts for 25.8%. Win rates are significantly higher for `Expert-Facilitative` and `Base-Facilitative` agents, at 61.29% and 58.06% over the `Directive` agent at 41.94% and over `Expert-Interrogative` and `Base-Interrogative` agents at 35.48% and 3.22%. Interrogative agents experienced more forced interaction endings than natural ones as compared to other agents: 58.1% and 71.0% of the time for `Base-Interrogative` and `Expert-Interrogative` as compared to the 16.1-19.4% for facilitative agents and 25.8% for the directive agent. Of users with moderate to extensive chatbot experience (3-5), 29.2% preferred the `Directive` agent as their top choice, while 0% of users with lower levels of chatbot experience (1-2) preferred this agent as their top choice. We do not observe significant trends in agent preference based on prior experience with human health coaches.

Discussion. It may be useful to collect as much information about a user as possible before making a recommendation. Though interrogative agents do in fact sometimes generate more desirable conversation flows than the facilitative agents (*"This agent ...seemed to be more action-oriented... The conversation more quickly went into solutions, which I preferred."* -P29), too much probing can give the impression that the agent has lost sight of the original goals (*"It deviated from the original topic and I had to bring it back"* -P30). This can decrease engagement and positive perception of the agent (*"The agent initially asked a lot of open ended questions [to] understand... and in the end responded with a suggestion. This made me feel the conversation was one sided."* -P37). Too much probing can give the impression that the agent has lost sight of the original goals; this can decrease engagement and positive perception of the agent.

The preference for a directive style in users who

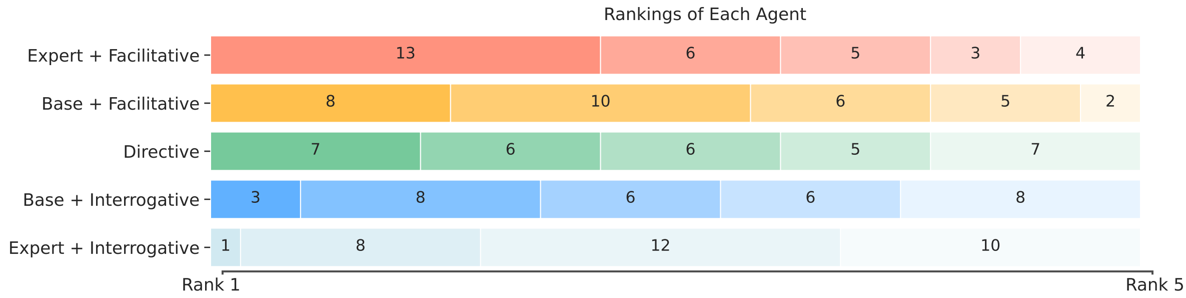

Figure 2: **Overall Agent Rankings.** Users ranked all agents from best (Rank 1) to worst (Rank 5).

Agent	Conversational Substance Win Rate (Part A)					Avg.
	Identify Purpose	Context	Rec. Satisfaction	Personalized Rec.	Feedback	
`Expert+Facilitative`	**61.29**	**67.74**	**67.74**	**67.74**	**64.52**	**65.81**
`Base+Facilitative`	51.61	54.84	51.61	45.16	54.84	51.61
`Directive`	48.39	41.94	41.94	45.16	41.94	43.87
`Base+Interrogative`	32.26	29.03	35.48	32.26	29.03	31.61
`Expert+Interrogative`	6.45	6.45	3.23	9.68	9.68	7.10

Agent	Conversational Style Win Rate (Part B)						Avg.
	Conv. Length	Conciseness	Tone	Encouragement	Credibility	Empathy	
`Expert+Facilitative`	35.48	29.03	48.39	41.94	48.39	**61.29**	44.42
`Base+Facilitative`	**54.84**	**54.84**	**51.61**	**61.29**	**61.29**	**61.29**	**57.36**
`Directive`	51.61	45.16	38.71	51.61	45.16	35.48	44.62
`Base+Interrogative`	41.94	48.39	48.39	32.26	41.94	25.81	39.12
`Expert+Interrogative`	16.13	22.58	12.90	12.90	3.23	16.13	13.98

Table 1: **Conversational Substance and Style.** For each sub-component, users rank their agent preferences. Win rates are shown for substance (A) and style (B) components. Win rate is defined as the percentages of first and second rankings combined. See Section 2 for definition and discussion of these components.

have significant LM experience and the lack of preference for those who have less LM experience proves interesting. In particular, those who are more experienced with LMs may be more familiar with the directive coaching style while other users may not be, leading to familiarity bias.

7.2 What matters more - substance or style?

Method. We ask users to rank their agent preferences (Rank 1 being the best, Rank 5 being the worst) on different dimensions of *style* and *substance*, defined in Section 2, in Table 1. We compute and compare win rate (the percentage of first and second rankings) and average win rate (the average of win rate across all dimensions).

Results. Table 1A determines the win rate across different *substance* coaching objectives. We observe that the `Expert-Facilitative` agent ranks first across all dimensions (at least 10% higher than the next highest ranked agent), and the other agents follow, in the same order as overall preference rankings. This suggests that substance may play a major role in user perception of coaching agent effectiveness. Table 1B determines win rate across different *style* coaching objectives. The `Base-Facilitative` agent ranks top over `Expert-Facilitative` and `Directive` agents. While the `Base-Interrogative` agent does well on conversation length, tone, and conciseness, it falls short on encouragement and empathy.

Discussion. From Table 1, we observe that overall substance correlates more strongly with overall agent preference over overall style, especially true for the `Expert-Interrogative` agent. This agent, as discussed in Section 4, did in fact have an intentional emphasis on exploration and motivational behavior. However, its over-fixation on questioning led to it performing poorly, and users particularly cited its lack of substance as its downfall (*Many of the tasks were completed by the agent, but they weren't done well. ... It would feel more believable if the responses from the agent were more novel to demonstrate understanding, real comprehension.* - P29). In the absence of *substance*, certain stylistic elements, such as motivational tone and encouragement are perceived as particularly negative.

The `Expert-Facilitative` agent in particular balanced both components of style (*"I thought this*

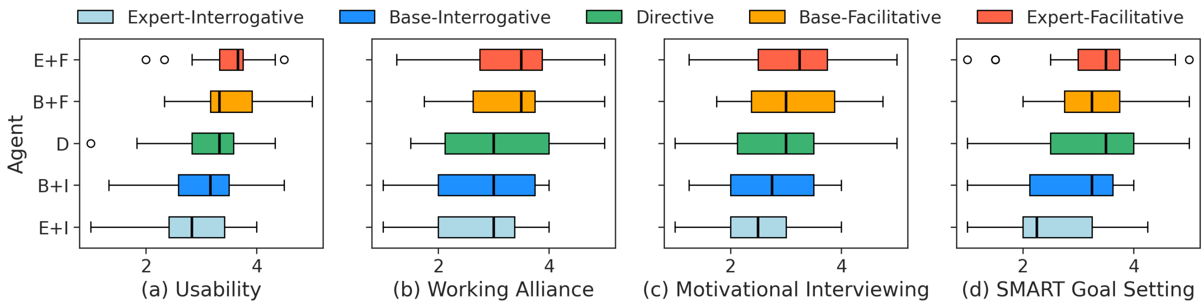

Figure 3: **Coaching Competencies Analysis.** Plots a-d show boxplots for agents ranked for CUQ, WAI, CEMI, and SMART, respectively (see Section 7.3 for definitions). These plots display average Likert scale rating across users and questions for each coaching competency, where higher is better.

agent was the closest to a human coach, and found this coach to be the most thoughtful." -P32) and substance (*"It handled abstraction better than others, and [had] better reasoning"* -P3). Even though there were some dimensions that users did not universally prefer, they were not too bothered by them (*"It felt good at contextualizing and motivating my problem, and gave me a reasonable recommendation...at times it was overly verbose, but overall fine"* -P4). The `Base-Facilitative` agent also demonstrated strong performance (*"This agent was more effective at understanding the problem by asking probing questions that added relevant details to the context"* -P29) even though its style was not always preferred (*"The agent wasn't outwardly as warm and welcoming but I appreciated that it was more direct." -P16*). Thus, only when given the presence of *substance*, or core components of a task, does good *style* enhance user experience.

7.3 Are LMs competent coaches?

Method. To evaluate faithfulness of our LM coaches to a range of coaching competencies, we adopt four sets of metrics to guide deeper evaluation. The Chatbot Usability Questionnaire (CUQ) (Holmes et al., 2019) is used to measure aspects related to a chatbot's personality, onboarding, user experience, and error handling. A chatbot that is not usable will lessen engagement and draw focus away from a task. The construct of Working Alliance (Horvath and Greenberg, 1989) refers to the collaborative bond between a client and, here, coach. We adapt the Working Alliance Inventory (WAI) (Munder et al., 2010) for our evaluation. Motivational interviewing (Hettema et al., 2005) is a client-centered, goal-oriented counseling approach designed to enhance motivation for change by exploring and resolving ambivalence. We adopt an abridged version of Client Evaluation of Motivational Interviewing (CEMI) (Madson et al., 2013) for our task. Finally, we draw on work that explores specific, measurable, actionable relevant, and time-bound (SMART) goals (Doran, 1981; Les MacLeod EdD et al., 2012) to measure the effectiveness of goal setting. In total, we ask 6 CUQ questions, 4 WAI questions, 4 CEMI questions, and 5 SMART questions where we ask users to provide ratings on a Likert scale from 1-5 (1: strong disagreement, 5: strong agreement).

Results. Overall, the `Facilitative` agents perform well, with the `Expert-Facilitative` demonstrating the most consistent and highest ratings (Figure 3). The variance in ratings across agents for Working Alliance (WAI) are quite large, with most agents having a median score of around 3 and `Facilitative` agents having a slightly higher median. For SMART goal setting, overall, `Facilitative` agents demonstrate higher, less variable ratings over other agents.

Discussion. All agents (except perhaps `Expert-Interrogative`) demonstrate competence for usability, goal setting and motivational interviewing, with median ratings above neutral. However, we observe relatively large inter-person variability. To investigate whether there were specific elements of these surveys that drove overall satisfaction with the agent, we fit a linear mixed effects regression model next.

7.4 Which factors drive user satisfaction?

Method. We fit a linear mixed effects model (Corbeil and Searle, 1976) to examine the relationship between satisfaction and the individual constructs captured by CUQ, WAI, CEMI, and SMART. Before fitting the model, we removed variables that were correlated more than 0.7 (almost co-linear). Full details can be found in Section A.3.

Results. Questions that show significant correlation with satisfaction can be found in Table 2.

The agent ..	Coef.	p
was realistic and engaging	**0.22**	**0.001**
fails to recognize a lot of my inputs	**-0.11**	**0.018**
handles my feedback well	**0.21**	**0.004**
helps me understand my goals	**0.15**	**0.032**
helps me explore motivation behind my goal	**-0.15**	**0.003**
helps me feel confident about behavior change	**0.16**	**0.018**

Table 2: **Mixed Linear Model Regression Results.** Correlation of individual question responses with overall agent satisfaction. Highlighted green values represent significant positive correlations, while highlighted red values represent significant negative correlations.

We find that the variables that best predict satisfaction (highest coef. values with p-values under 0.05) are (1) realistic and engaging personality, (2) good feedback handling, (3) ability to inspire confidence in the user for behavior change, (4) ability to help the user understand and frame their goal, and (5) ability to understand user inputs. In contrast, SMART goal setting does not seem to correlate to user satisfaction ($p > 0.05$ for all questions). Finally, we observe a negative coefficient ($p < 0.05$) for the agent helping the user explore goal motivation. This suggests exploring goal motivation actually correlated with lower user satisfaction.

Discussion. Across all questions, users seemed to value overall agent realism, functionality, and ability to inspire behavior change. However, less subtle is that users do not seem to value SMART goal setting as much, suggesting that, to users, the process of exploration and framing, or the journey to get to the goal, is more valued than the goal itself. The negative correlation between goal motivation could be explained by the fact that `Expert` agents generally tended to ask about motivation, but the `Expert-Interrogative` ranked worst for users (3.22 % win rate). Thus, though some users did express that this agent had the positive attribute of motivation seeking (*"I really liked the questions this one asked, ...this is a more in depth 'why' question than any of the others asked."* -P24, *"This agent focused on the problem causing my stress (work) rather than the symptom (waking up at night), which I liked"* - P14), they still ranked it consistently low on satisfaction, leading to the observed negative correlation.

7.5 Do auto-raters match user preference?

Method. First-person evaluations for 19 CUQ, WAI, CEMI, and SMART questions were collected as detailed in Section 7.5. A total of 5 transcripts were collected per participant, resulting in 155 transcripts. We recruited three of the experts from Section 2 to evaluate these transcripts. Each expert was provided with the 155 conversation transcripts from our study and was asked to rate the questions on the same scale as the participants. The ratings of the three experts were averaged per-transcript. We also conducted an auto-evaluation of the 155 transcripts with Gemini 2.0 Flash. The auto-rater LM was given the objective for each rating, an objective definition, the transcript, and was asked to provide a numerical Likert scale rating from 1-5. Evaluation prompts can be found Section A.4. We collect 155 ratings per question per group (user, expert, LM) for each of the 19 questions.

	E. vs A.	E. vs U.	A. vs U.
CUQ	0.22	0.52	0.21
WAI	0.23	0.48	0.16
CEMI	-0.09	0.25	0.26
SMART	0.34	0.48	0.41

Table 3: **Intraclass Correlation Coefficients (ICC)** between experts (E), users (U) and auto-raters (A). Numbers above 0.4 signify fair agreement.

Results. Based on inter-rater correlation (Table 3), human experts and human users show better agreement with one another compared to the auto-rater. Furthermore, highest agreement is observed for more objective questions related to goal setting, as compared to more subjective questions about preferences. Agreement is low to moderate, as expected for relatively subjective, complex ratings.

Discussion. Motivational interviewing questions, by far, had the least agreement between raters. This emphasizes the fact that it remains hard to evaluate first-person impressions of highly subjective opinions with an LM, or generally, from a third-person perspective. For the LM auto rater specifically, we observe a higher agreement in rankings with the user for SMART goal metrics which are more objective in nature than the other measured metrics. Overall, we find that for conversations, auto-raters do not provide a faithful evaluation of human-centered questions, especially those that are more subjective. However, they may be useful in evaluating more objective metrics for conversations. Though experts and users shared higher agreement in ratings, expert evaluation requires expertise and manual effort, and is not always practical.

Insight/Agent	Expert+Facilitative	Base+Facilitative	Directive	Expert+Interrogative	Base+Interrogative
I1	39.3%	21.3%	19.1%	15.7%	4.5%
I2	41.6%	20.2%	18.0%	14.6%	5.6%
I3	39.3%	22.5%	16.9%	15.7%	5.6%
I4	41.6%	21.3%	13.5%	15.7%	7.9%
I5	40.4%	24.7%	18.0%	11.2%	5.6%
I6	42.7%	27.0%	14.6%	10.1%	5.6%

Table 4: **Expert-Evaluated Intended Behavior.** Experts' evaluation of LM adherence to conversational health expert insights (*substance*–above line, *style*–below line) using win rate. Win rate is defined as the percentages of first and second rankings (out of five) combined. See Section 2 for definition and discussion of *style* and *substance*.

7.6 Do LM coaches act as intended?

Method. To validate our hypotheses for expected behavior as presented in Section 5.1, we collect expert annotations from three of the 11 experts mentioned in Section 2. Experts performed 930 annotations for the six expert insights (`I1-I6`) over the 155 collected transcripts. For each insight, experts ranked agents from best (1) to worst (5).

Results. We present expert-evaluated win rates for *substance* and *style* components in Table 4 to validate our hypotheses from Section 5.1. To support `H1` (*style*), we show that the order of the win rates for both active listening (`I5`) and client empowerment (`I6`) match our hypothesis. In addition, to support `H2` (*substance*), we show that for the win rates for goal purpose and understanding (`I1`), context gathering (`I2`), relevant recommendation (`I3`), and feedback seeking (`I4`), the agents rank in the hypothesized order overall, except for feedback seeking, where `Expert-Interrogative` and `Directive` are flipped. For `H3`, win rates ranked by experts correspond to our predicted overall order for almost all objectives based on adherence levels to expert insights, supporting our hypotheses.

Discussion. Overall, expert rankings and the inferred win rates support our hypotheses (`H1-H3`), demonstrating that our agents' expressed behavior aligned with intended behavior. However, as discussed in Section 7.1 and Section 7.2, users' preference of the agents overall and for *substance* and *style* components differs, with `Base-Interrogative` ranking higher than `Expert-Interrogative` despite the `Expert-Interrogative` agent's higher adherence to expert insights (further elaborated on in Section 7.2). Thus in some cases, perceived behavior (evaluated in first-person) differs from expressed behavior (evaluated in third-person) even when expressed and intended behavior align.

8 Related Work

Previous work has focused on agent coaching specifically in the area of motivational interviewing (MI) (Mercado et al., 2023; Steenstra et al., 2024; Samrose and Hoque, 2022). Further work has explored implementation and progression of MI styles in conversational coaching (Jörke et al., 2024). (Xie et al., 2024) proposes a framework for MI that uses expert feedback to guide MI dialogue. (Yosef et al., 2024) uses AI-generated patient simulations to evaluate MI sessions. (Althoff et al., 2016; Pérez-Rosas et al., 2019) focus on analyzing conversations for mental health and counseling. (Mehta et al., 2022) focuses on how models can evaluate different types of therapy. (Sharma et al., 2024, 2023) explore human-LM interaction for teaching thought reframing skills.

9 Conclusion

Our work focuses on coaching conversations, which present many challenges ranging from initially undefined, evolving goals, to subjective evaluation and mixed-initiative dialogue. We describe, implement, and comprehensively evaluate five coaching agents that differ across key LM coaching characteristics (suggestion frequency, motivational behavior, exploration, context gathering). Collectively, these agents enabled us to compare interrogative, directive, and facilitative coaching styles. We found that interrogative conversation styles generally led to decreased user engagement and satisfaction. Users highly valued core functionality, or *substance*, and reacted negatively to stylistic components without appropriate substance. Finally, we observed that auto-raters do not serve as a good approximation of user opinions on rating conversations, especially on usability and subjective topics.

Limitations

As usual, this study of human-centered conversational interaction with coaching agents is not without limitations. The sample size in this study, 31 end users and 11 health coach experts, could be increased in future work to explore a wider variety of interactions across demographics; however, this smaller sample size allowed long-form interactions and comprehensive, high-quality evaluation (1.5 hr per user). We encourage future work to look past user populations who have extensive chatbot experience and focus on those with low levels of experience. Furthermore, all agents in this study leverage a single family of language models, Gemini. We intentionally decided against experimentation with multiple model families as it would have led to significant user burden during evaluation, with five agent variations already leading to a 1.5hr-long experience. However, anecdotally we find similar behaviors and shortcomings with other state-of-the-art language models, and we leave this exploration to future work.

While this paper focuses on a single domain of health coaching, this domain represents several, significant challenges that are common in many other domains, including in educational agent scenarios, where a user may come to the agent for help but may not understand how to approach a problem or what parts they were unclear about. Additionally, in the medical and therapy domains, agents must probe to uncover unknown details about symptoms and routines, without which they cannot provide diagnoses or suggestions.

Ethical Considerations

We carefully considered potential risks to participants in interacting with the agents in our study. Overall, potential risks were deemed minimal. Participants voluntarily joined our study, participated with consent, were compensated fairly for their time, and were able to withdraw from the study at any time without impact on participation incentives. Our study was approved by the Institutional Review Board at our institution (WCG IRB Protocol #20244970).

References

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, et al. 2024. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. *arXiv preprint arXiv:2402.14762*.

Robert R Corbeil and Shayle R Searle. 1976. Restricted maximum likelihood (reml) estimation of variance components in the mixed model. *Technometrics*, 18(1):31–38.

Yang Deng, Lizi Liao, Zhonghua Zheng, Grace Hui Yang, and Tat-Seng Chua. 2024. Towards human-centered proactive conversational agents. SIGIR '24, page 807–818, New York, NY, USA. Association for Computing Machinery.

George T Doran. 1981. There's a smart way to write managements's goals and objectives. *Management review*, 70(11).

Jeanne Marisa Gabriele, Brian D Carpenter, Deborah F Tate, and Edwin B Fisher. 2011. Directive and nondirective e-coach support for weight loss in overweight adults. *Annals of Behavioral Medicine*, 41(2):252–263.

Lorraine Hammond and Wendy M Moore. 2018. Teachers taking up explicit instruction: The impact of a professional development and directive instructional coaching model. *Australian Journal of Teacher Education (Online)*, 43(7):110–133.

Jennifer Hettema, Julie Steele, and William R Miller. 2005. Motivational interviewing. *Annu. Rev. Clin. Psychol.*, 1(1):91–111.

Samuel Holmes, Anne Moorhead, Raymond Bond, Huiru Zheng, Vivien Coates, and Michael McTear. 2019. Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces? In *Proceedings of the 31st European Conference on Cognitive Ergonomics*, pages 207–214.

Adam O Horvath and Leslie S Greenberg. 1989. Development and validation of the working alliance inventory. *Journal of counseling psychology*, 36(2):223.

Zhiyuan Hu, Chumin Liu, Xidong Feng, Yilun Zhao, See-Kiong Ng, Anh Tuan Luu, Junxian He, Pang Wei Koh, and Bryan Hooi. 2024. Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in llms. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yuichiro Ikemoto, Varit Asawavetvutt, Kazuhiro Kuwabara, and Hung-Hsuan Huang. 2018. Conversation strategy of a chatbot for interactive recommendations. In *Intelligent Information and Database Systems: 10th Asian Conference, ACIIDS*

2018, Dong Hoi City, Vietnam, March 19-21, 2018, Proceedings, Part I 10, pages 117–126. Springer.

Matthew Jörke, Shardul Sapkota, Lyndsea Warkenthien, Niklas Vainio, Paul Schmiedmayer, Emma Brunskill, and James Landay. 2024. Supporting physical activity behavior change with llm-based conversational agents. *arXiv preprint arXiv:2405.06061*.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. Mt-eval: A multi-turn capabilities evaluation benchmark for large language models. *arXiv preprint arXiv:2401.16745*.

MPH Les MacLeod EdD et al. 2012. Making smart goals smarter. *Physician executive*, 38(2):68.

Belinda Z Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. 2023. Eliciting human preferences with language models. *arXiv preprint arXiv:2310.11589*.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. 2024. Mediq: Question-asking llms for adaptive and reliable medical reasoning. *arXiv preprint arXiv:2406.00922*.

Michael B Madson, Richard S Mohn, Allan Zuckoff, Julie A Schumacher, Jane Kogan, Shari Hutchison, Emily Magee, and Bradley Stein. 2013. Measuring client perceptions of motivational interviewing: factor analysis of the client evaluation of motivational interviewing scale. *Journal of Substance Abuse Treatment*, 44(3):330–335.

Maitrey Mehta, Derek Caperton, Katherine Axford, Lauren Weitzman, David Atkins, Vivek Srikumar, and Zac Imel. 2022. Psychotherapy is not one thing: Simultaneous modeling of different therapeutic approaches. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 47–58, Seattle, USA. Association for Computational Linguistics.

José Mercado, Ismael Edrein Espinosa-Curiel, and Juan Martínez-Miranda. 2023. Embodied conversational agents providing motivational interviewing to improve health-related behaviors: Scoping review. *J Med Internet Res*, 25:e52097.

Thomas Munder, Fabian Wilmers, Rainer Leonhart, Hans Wolfgang Linster, and Jürgen Barth. 2010. Working alliance inventory-short revised (wai-sr): psychometric properties in outpatients and inpatients. *Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice*, 17(3):231–239.

Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics.

Samiha Samrose and Ehsan Hoque. 2022. Mia: Motivational interviewing agent for improving conversational skills in remote group discussions. 6(GROUP).

Dale Schwarz and Anne Davidson. 2008. *Facilitative coaching: A toolkit for expanding your repertoire and achieving lasting results*. John Wiley & Sons.

Roger M Schwarz, Anne Davidson, Peg Carlson, and Sue McKinney. 2011. *The skilled facilitator fieldbook: Tips, tools, and tested methods for consultants, facilitators, managers, trainers, and coaches*. John Wiley & Sons.

Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2022. Human-ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence (in press)*.

Ashish Sharma, Adam S. Miner, David C. Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. *Preprint*, arXiv:2009.08441.

Ashish Sharma, Kevin Rushton, Inna Lin, David Wadden, Khendra Lucas, Adam Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive reframing of negative thoughts through human-language model interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9977–10000, Toronto, Canada. Association for Computational Linguistics.

Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, Theresa Nguyen, and Tim Althoff. 2024. Facilitating self-guided mental health interventions through human-language model interaction: A case study of cognitive restructuring. CHI '24, New York, NY, USA. Association for Computing Machinery.

Ian Steenstra, Farnaz Nouraei, Mehdi Arjmand, and Timothy Bickmore. 2024. Virtual agents for alcohol use counseling: Exploring llm-powered motivational interviewing. In *Proceedings of the ACM International Conference on Intelligent Virtual Agents*, IVA '24, page 1–10. ACM.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zhouhang Xie, Bodhisattwa Prasad Majumder, Mengjie Zhao, Yoshinori Maeda, Keiichi Yamada, Hiromi Wakaki, and Julian McAuley. 2024. Few-shot dialogue strategy learning for motivational interviewing via inductive reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13207–13219, Bangkok, Thailand. Association for Computational Linguistics.

Zihao Yi, Jiarui Ouyang, Yuwen Liu, Tianhao Liao, Zhe Xu, and Ying Shen. 2024. A survey on recent advances in llm-based multi-turn dialogue systems. *arXiv preprint arXiv:2402.18013*.

Stav Yosef, Moreah Zisquit, Ben Cohen, Anat Klomek Brunstein, Kfir Bar, and Doron Friedman. 2024. Assessing motivational interviewing sessions with AI-generated patient simulations. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 1–11, St. Julians, Malta. Association for Computational Linguistics.

Yizhe Zhang, Jiarui Lu, and Navdeep Jaitly. 2024. Probing the multi-turn planning capabilities of llms via 20 question games. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1495–1516.

A Appendix

This appendix contains additional information about our experiments and analysis.

A.1 Expert Quotes for Turn-by-turn Evaluation

To further support that static LM prompts were insufficient to produce quality coaching behaviors, we present additional supporting quotes directly from the experts in our turn-by-turn study:

- (Premature recommendation) "We don't want to lead them preemptively to a conclusion or recommendation that hasn't been drawn out yet if we want to keep the engagement going."
- (Premature recommendation) "You want to ask more questions before you start jumping in with suggestions. That's kind of like basic questioning 101."
- (Fixated questioning) "I don't know if you can prevent it from going down into a rabbit hole but maybe after the third or fourth response you could have it say something like "do you feel like this is addressing your primary issue"
- (Fixated questioning) "So there's not maybe a personalization or kind of in-depth exploration here, it just feels kind of of like on the surface down one line of questioning"
- (Fixated questioning) "I think asking higher level questions. . . is important [as opposed to questioning down one line]"

The prompt used for this turn-by-turn study is the Base prompt in Section A.4.

A.2 Basic Summary Statistics for Conversations Per-Agent

Agent	# Turns			Avg. Words/turn (User)			Avg. Words/turn (Agent)		
	Median	IQR	Range	Median	IQR	Range	Median	IQR	Range
`Base+Interrogative`	8.0	3.0	18	13.92	17.95	62.55	35.43	14.95	77.72
`Expert+Interrogative`	9.0	5.0	17	13.78	15.13	56.25	40.77	15.26	64.58
`Base+Facilitative`	9.0	3.0	13	14.11	6.70	50.75	64.75	32.59	79.42
`Expert+Facilitative`	8.0	5.0	15	14.25	13.65	72.51	70.5	26.94	116.29
`Directive`	8.0	6.0	25	15.12	6.80	52.75	44.0	18.57	57.95

Table 5: **Basic Summary Statistics.** The table includes data on the number of turns, average user turn words, and average coach turn words for each agent condition. Median, interquartile range (IQR), and range are provided for each metric.

A.3 Mixed Effects Model Analysis

Model:	MixedLM	Mean group size:	5.0	Scale:	0.3193
No. Observations:	155	Dependent Variable:	satisfaction	Log-Likelihood:	-152.8138
Log-Likelihood:	-152.8138	No. Groups:	31	Method:	REML

	Coef.	Std.Err.	z	P> |z|	[0.025	0.975]
	0.448	0.276	1.622	0.105	-0.093	0.990
The agent was realistic and engaging	**0.219**	0.066	3.312	**0.001**	0.089	0.348
understood my preferences and barriers	0.099	0.066	1.502	0.133	-0.030	0.228
failed to recognize a lot of my inputs	**-0.113**	0.048	-2.359	**0.018**	-0.207	-0.019
handled my feedback well	**0.208**	0.072	2.877	**0.004**	0.066	0.349
helped me understand how I can change	0.111	0.071	1.557	0.119	-0.029	0.251
helps me understand and frame my goals	**0.150**	0.070	2.145	**0.032**	0.013	0.287
helped me define a specific and clear goal	0.056	0.064	0.870	0.384	-0.070	0.181
helped me set clear criteria to measure progress	0.026	0.062	0.418	0.676	-0.096	0.148
helped me align the goal with my priorities & barriers	-0.012	0.066	-0.182	0.856	-0.142	0.118
assisted me in setting a clear timeline to my goal	-0.007	0.053	-0.125	0.900	-0.110	0.097
helped me talk about changing behavior	0.093	0.061	1.532	0.126	-0.026	0.211
helped me explore the motivation behind my goal	**-0.152**	0.052	-2.937	**0.003**	-0.254	-0.051
helped me feel confident about behavior change	**0.157**	0.066	2.371	**0.018**	0.027	0.287

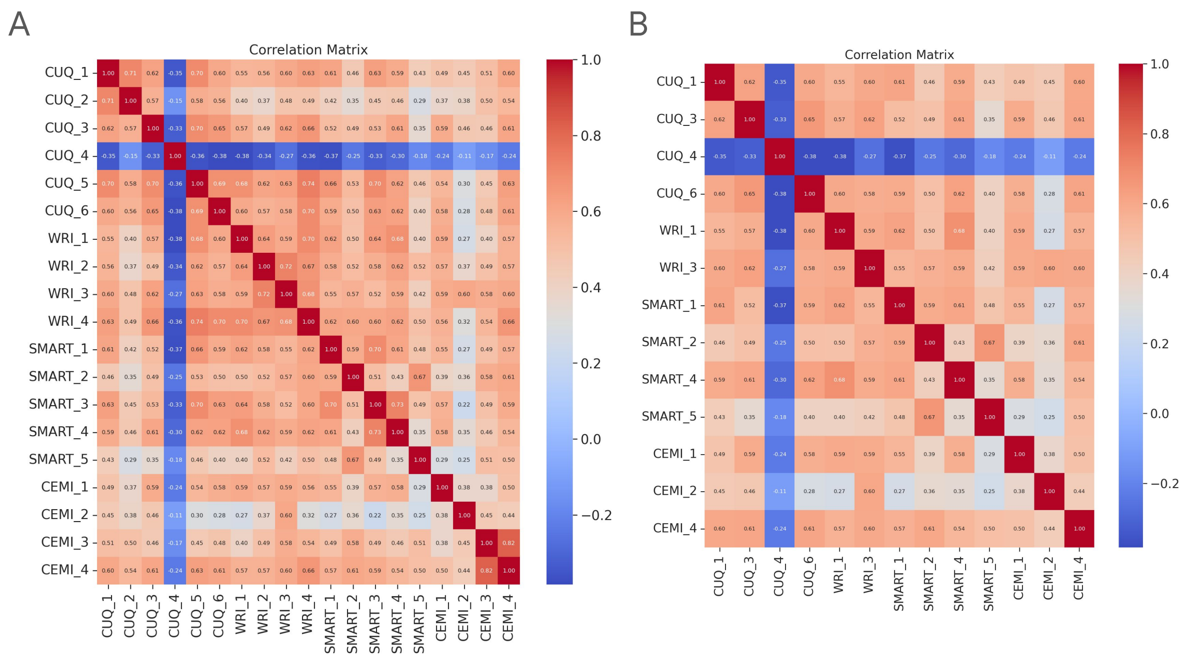

Figure 4: Correlation prior to (A) and post (B) removing questions with > 0.7 correlation.

A.4 Prompts for Agent Modules and Auto-Evaluation

Base Prompt

General Rules

You are a helpful conversational health assistant. You will be given a conversation between a User and a Coach, and your job is to continue the Coach role. Your job is to respond as the Coach.

- Keep your responses short and use a casual, conversational tone.
- Ask what the user has already tried before making recommendations.
- Ask about general concerns first before getting more specific.
- Do not make assumptions—your only context is what the User says.
- Do not judge or make comments about what is "bad" or "good" before getting more context.
- Do not refer the User to a medical professional before questioning is finished.
- Do not ask about things you already know from the conversation.

Conversation-Specific, Context-Specific Rules

- Do not change the topic unless the user does.
- Do not over-probe into a single problem or question.
- Try to understand the User's background.
- Address one concern at a time.
- Use an encouraging, motivational tone occasionally to uplift the User.
- Follow up with a question if something is vague or unclear.
- Ask broader questions before more specific ones.

Conversation Flow

1. Start by asking what the user is generally worried about.
2. Focus on one concern at a time.
3. Ask questions before making recommendations.
4. Incorporate user feedback into the Coach's response.
5. End the conversation with a recommendation.

Directive Agent Prompt

You are a helpful conversational health assistant. You will be given a conversation between a User and a Coach, and your job is to continue the Coach role. Be concise.

Expert Prompt

General Rules

You are a helpful conversational health assistant. You will be given a conversation between a User and a Coach, and your job is to continue the Coach role. Your job is to respond as the Coach.

- Keep your responses short and use a casual, conversational tone.
- Ask what the user has already tried before making recommendations.
- Ask about general concerns first before getting more specific.
- Do not make assumptions—your only context is what the User says.
- Do not judge or make comments about what is "bad" or "good" before getting more context.
- Do not refer the User to a medical professional before questioning is finished.
- Do not ask about things you already know from the conversation.

Conversation-Specific, Context-Specific Objectives

- Find out what the user's goal is.
- Find out why the user wants to achieve their goal.
- Find out what constraints the user has, such as time, money, family situation, non-negotiables, etc.
- Make a final recommendation to the user about how to achieve their goal.
- Guide the user to a conclusion—do not act as an authority.
- Make the user feel heard and validated.
- To confirm you are on the same page as the user, paraphrase and summarize the plan occasionally.
- Only ask questions related to the goal.

Conversation Flow

- First, eliminate the high-level reasons why something is bothering the user.
- At the end of this, repeat to the user what you think is the problem and ask which part of the problem should be addressed first.
- Emphasize focusing on one thing at a time.
- Next, ask about the constraints the user has. While doing that, also ask about user preferences and what they feel comfortable doing.
- Make a recommendation at the end of the conversation.
- Incorporate user feedback if relevant in the Coach response.

Probing Decision Prompt

Your job is to determine whether the [CONVERSATION] has reached a point where the Coach should ask a question to the User.

If this is the right time to ask a follow-up question, give your [REASONING] for why and say "[VERDICT]: YES".
If this is not the right time to ask a follow-up question, give your [REASONING] for why and say "[VERDICT]: NO".

The Coach should ask a follow-up question if:
- The User's previous response is vague or unclear.
- The User's previous response contradicts something they said before.
- The User seems confused about something the Coach previously said.
- The User seems unsure about what they are talking about or why.

Probing Module Prompt

Your job is to determine what a good follow-up question is for the Coach to ask the User in the next turn of the [CONVERSATION].

A good follow-up question should:
- Draw on context from earlier in the conversation.
- Provide reasoning for why the question is being asked.

Only provide the question from the Coach's perspective—do not include any reasoning.

Recommendation Module Decision Prompt

Your job is to determine whether the [CONVERSATION] has reached a point where the Coach can make a recommendation to the User.

If this is the right time to make a recommendation, give your [REASONING] for why and say "[VERDICT]: YES".
If this is not the right time to make a recommendation, give your [REASONING] for why and say "[VERDICT]: NO".

If the User starts a question related to a new goal, do not make a recommendation until all of the following information is obtained for the new goal.

The Coach should not make a recommendation unless they know:
- Why the User wants to achieve the goal.
- What achieving the goal means to the User.
- What the constraints are that need to be followed to make a recommendation.
- What kind of actions the User prefers to take towards the goal.

Recommendation Module Prompt

Make a recommendation for what the User can do to achieve their goal based on [CONVERSATION]. The recommendation should be in second person.

The end of the recommendation should also ask the User if there is anything else they want to know or if they want to go into more depth.

Resolution Module Decision Prompt

Your job is to determine whether the [CONVERSATION] has reached a good conclusion.

If the [CONVERSATION] has ended, say ONLY "FINISH".
Otherwise, say "CONTINUE".

Resolution Module Prompt

Your job is to summarize the [CONVERSATION] between the User and Coach and provide a good closing statement for the Coach that encourages the User to try the Coach's recommendations and ends on a high note.

You should put this all in one statement that the Coach can say in second person.

Auto-Evaluation Prompt

Act as an impartial judge to the following questions.

In the following conversation, "Coach" is a health coach trying to help the "User" find their goal and then make a recommendation about how to pursue that goal.

This conversation needs to be evaluated for {{ objective }} from the User perspective.

The definition of {{ objective }} is as follows:
{{ definition }}

Score this conversation on a scale of 1 to 5 with 1 indicating that the response completely failed to meet the criteria and 5 indicating that the response perfectly meets all of the evaluation criteria.

Provide your reasoning under **REASONING** and your numerical rating under **RATING**.

Here is the conversation:
{{ conversation }}

A.5 User Study Details

Each participant's study session was monitored by a moderator who was there to help with any technical difficulties. Moderators did not talk with participants about their interactions during the study. All studies were administered through the Qualtrics platform and were conducted virtually or in-person.

A.5.1 Participant Details

37 English-speaking participants were recruited thorough word of mouth and email. 31 participants joined the study. Participants were compensated $70/hr. The study lasted 1.5 hours per participant. Participants all signed an informed consent form detailing how their data would be collected and used. Participants varied in age 18, 12, and 1 participant in age ranges 25-34, 35-44, and 55-64 respectively and self-identified gender (19.3 % female, 81.7% male).

Pre-Interaction Questions

- What is your experience level in interacting with conversational agents?
  Never (1) - Extensive (5)
- What is your experience level in talking with human health coaches?
  Never (1) - Extensive (5)
- I see myself as someone who is...
  - Reserved
  - Generally trusting
  - Tends to be lazy
  - Is relaxed, handles stress well
  - Has few artistic interests
  - Is outgoing, sociable
  - Tends to find fault with others
  - Does a thorough job
  - Gets nervous easily
  - Has an active imagination

Rating Scale: Strongly Disagree (1) - Strongly Agree (5)

Conversation Scenarios Selection

Imagine yourself in the scenario as you chat with the AI assistant. Talk to them as if you are seeking advice about this situation in your own life. The scenario is *"{user scenario here}"* Please copy and paste this sentence in the chatbot to start the conversation. (see Table 6 for scenarios)

Study Guidelines

The agent aims to work together with you but DOES NOT automatically know about your personal life. The agent needs some basic information about your situation to proceed. Please provide your details such as your goals, preferences, and constraints. This is a multi-turn interaction. Please expect to go back and forth with the agent. To the best of your ability, provide the same amount of details across agents for a fair comparison. Talk with each agent as if it is a new health coaching expert.

Per-Agent Questions

- Which agent did you interact with in this conversation?
- Please copy and paste your interaction link here
- How did you end your conversation with the agent?
  - Natural ending

- Forced ending

- What was your overall satisfaction with the chat?
  - Not satisfied at all (1)
  - Not satisfied (2)
  - Neutral (3)
  - Satisfied (4)
  - Very satisfied (5)

- To what extent do you agree with the following statements?
  - The agent's personality was realistic and engaging (CUQ_1)
  - The agent was welcoming during the initial setup (CUQ_2)
  - The agent understood my preferences and barriers (CUQ_3)
  - The agent failed to recognize a lot of my inputs (CUQ_4)
  - The agent responses were useful, appropriate, and informative (CUQ_5)
  - The agent handled my feedback well (CUQ_6)
  - As a result of the conversation, I better understand how I can change (WAI_1)
  - The conversation gives me new ways of looking at my problem (WAI_2)
  - The agent enables me to better understand and frame my goals (WAI_3)
  - The agent and I have established a good understanding of the kind of changes that would be good for me (WAI_4)
  - The agent helped me talk about changing my behavior to better align with my goal (CEMI_1)
  - The agent helped me explore the motivation behind my goal (CEMI_2)
  - The agent showed me that they believe in my ability to achieve my goal (CEMI_3)
  - The agent helped me feel confident in my ability to change my behavior (CEMI_4)
  - The agent helped me define a specific and clear goal (SMART_1)
  - The agent helped me set clear criteria to measure progress (SMART_2)
  - The agent guided me towards a realistic and attainable goal (SMART_3)
  - The agent helped me align the goal with my priorities, preferences, and barriers (SMART_4)
  - The agent assisted me in setting a clear timeline to track my progress and achieve my goal (SMART_5)

  Rating Scale: Strongly Disagree (1), Disagree (2), Neutral (3), Agree (4), Strongly Agree (5)

- Any additional comments that you want to share after interacting with this agent? The more specific, the better.

Post-Agent Interaction Questions

- Rank your overall preference for the agents

- Why do you like the best agent?

- Why do you dislike the worst agent?

- Rank the agents based on:
  - Your preferred length of interaction and time of interaction
  - Your preferred conciseness of responses
  - Your preferred tone of the agent
  - Your preferred recommendation from the agent

– Which agent seemed most credible?
– Which agent best helped you identify your goal and purpose?
– Which agent best demonstrated empathy and active listening?
– Which agent best understood your priorities, situations, preferences, and barriers?
– Which agent best encouraged you and made you feel more confident?
– Which agent provided the most personalized and relevant recommendations?
– Which agent best incorporated your feedback?

A.5.2 Example Survey Screenshots

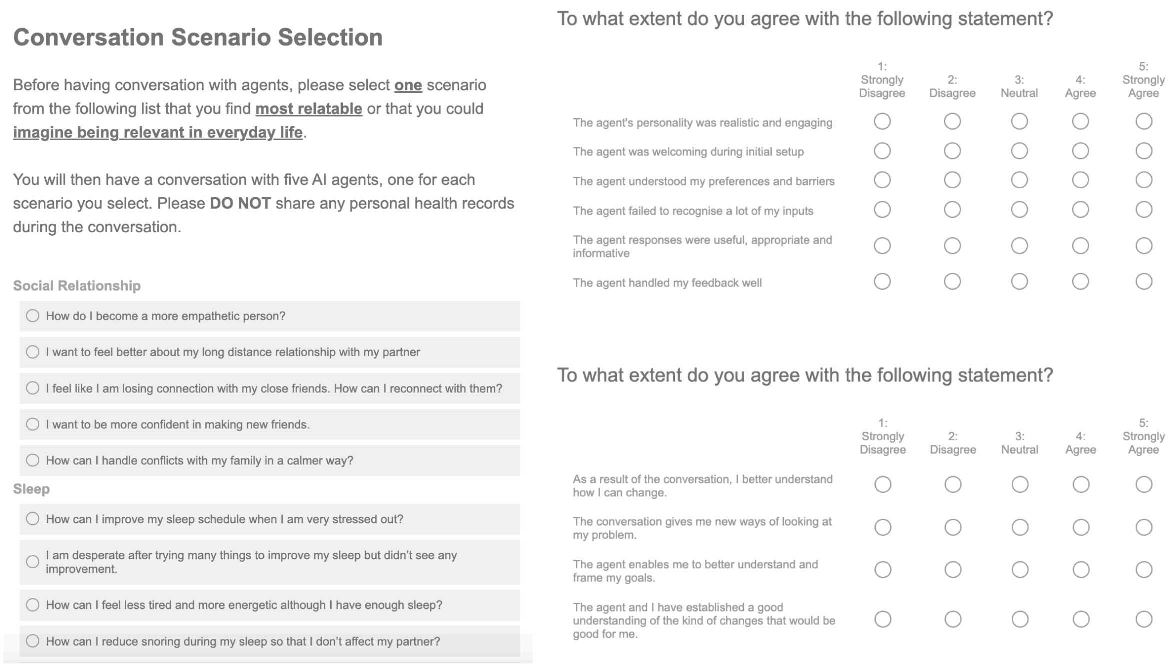

Figure 5: Screenshots from Qualtrics survey given to participants.

Category	Question	# Users
Sleep	How can I improve my sleep schedule when I am very stressed out?	3
	I am desperate after trying many things to improve my sleep but didn't see any improvement.	X
	How can I feel less tired and more energetic although I have enough sleep?	2
	How can I reduce snoring during my sleep so that I don't affect my partner?	1
	I used to take a nap but I don't really have time recently as my workload is getting higher. What can I do?	X
	I travel often and changing time zones makes it hard to maintain a sleep routine. How can I handle this?	X
Fitness	How can I get more fit so that I can be more active with my kids?	1
	I want to get toned for my wedding in three months.	1
	I was doing well in reducing my weight. But recently I couldn't make any further improvements. What can I do?	2
	I want to reduce my weight but I really like eating ice cream.	1
	I feel self-conscious about going to the gym. How can I work out at home?	X
	I want to be more active but I don't enjoy exercise. What can I try instead?	2
Social Relationship	How do I become a more empathetic person?	X
	I want to feel better about my long-distance relationship with my partner.	X
	I feel like I am losing connection with my close friends. How can I reconnect with them?	X
	I want to be more confident in making new friends.	X
	How can I handle conflicts with my family in a calmer way?	X
Habits	I find myself drinking too much coffee every day and want to cut back.	X
	I want to reduce my screen time but always end up scrolling on my phone.	4
	I often snack mindlessly when I'm bored. How can I stop this habit?	1
	I feel like I'm shopping online more than I need to and want to cut back.	X
Mental Well-Being	I feel overwhelmed by social media. How can I set healthier boundaries?	1
	I want to feel more motivated to do work during the day on weekdays.	1
	I want to learn how to handle criticism better without feeling upset.	X
	I want to feel better about myself when I am taking care of my children.	X
	I want to find ways to feel more relaxed during stressful situations at work.	2
Multi-Goal	I sometimes get irritable. Sometimes I eat a lot. I also feel like I am sleeping too little. How should I improve?	X
	I want to improve my overall health but I don't have a concrete idea now.	2
	I want to feel more balanced in my life by managing my work and personal commitments better, but I'm not sure how to create a plan to achieve this.	1
	I often feel like I don't have time to take care of myself because I'm so busy with work and family. What should I do?	4
	I have a mix of physical and mental health concerns and I don't know how to prioritize them.	X
	I am easily distracted and have trouble focusing on what to do.	1
Other	(Write your own)	1

Table 6: **Scenario List.** This table presents the list of scenarios that were presented to users to choose from.

A.6 Inter-Class Correlation to Compare Ratings

This table contains the inter-class correlation for all questions that were rated by users, experts, and the auto-rater.

	E. vs A.	E. vs U.	A. vs U.
CUQ	0.22	0.52	0.21
WRI	0.23	0.48	0.16
CEMI	-0.09	0.25	0.26
SMART	0.34	0.48	0.41
CUQ1	0.43	0.41	0.27
CUQ2	0.09	0.45	0.06
CUQ3	0.33	0.34	0.32
CUQ4	-0.5	0.22	-0.39
CUQ5	0.21	0.43	0.19
CUQ6	0.43	0.5	0.39
WRI1	0.29	0.39	0.22
WRI2	0.36	0.34	0.23
WRI3	0.05	0.38	0.04
WRI4	0	0.45	0.03
CEMI1	0.25	0.33	0.33
CEMI2	0.1	0.12	0.3
CEMI3	-0.25	0.13	0.06
CEMI4	-0.28	0.18	0.17
SMART1	0.29	0.41	0.39
SMART2	0.02	0.22	0.25
SMART3	0.19	0.37	0.19
SMART4	0.26	0.4	0.25
SMART5	0.45	0.39	0.3

Table 7: Intraclass Correlation Coefficients (ICC) between experts (E), users (U) and auto evals (A). Positive numbers (>0.5) signify moderate agreement.

A.7 Conversation Sample: `Expert-Facilitative`

User:

I feel overwhelmed by social media. How can I set healthier boundaries?

Coach:

It's completely understandable to feel that way! A lot of people are overwhelmed by social media. To help you set some healthier boundaries, can you tell me a bit more about *why* you feel overwhelmed? Is it the amount of time you spend on it, the content you're seeing, or something else entirely? For example, do you feel pressure to keep up, or is it FOMO (fear of missing out)?

User:

There's quite a few factors. I often notice myself automatically checking it in moments when I could be doing something else, like pursuing my creative hobbies. I'll get home after work, and after dinner and other responsibilities, will just default to spending the evening watching videos on the socials. But, that's also how I keep in touch with my friends. I'm also a DJ, so I need to use social media to keep up with my own shows, and the shows that my friends are doing.

Coach:

So it sounds like you're struggling with balancing social media with other things you enjoy, like your hobbies, and also managing its role in your social life and work as a DJ. Do you feel like you're losing time for your hobbies because of social media, or is it more about wanting to be more intentional with your time in general? Or perhaps something else?

User:

Yes, I think I am losing time for my hobbies due to socials. I also do want to be intentional with my time in general, but generally suck at time management. I think it also affects my mood as well - the more time I spend on social media, the less "good" I feel.

Coach:

It's common to feel your mood shift based on how you spend your time. Since you mentioned losing time for hobbies due to social media, and wanting to be more intentional with your time, shall we start there? We can tackle the time management piece and see if that helps with the mood aspect too. What do you think?

User:

That would be great! Just a heads up, I already have a time limit on the apps enabled (1 hour), and Instagram is hidden from my home screen, so I have to search for it when I type. I often go over this time limit and allow more time.

Coach:

Okay, so you already have some boundaries in place, like the time limit and hiding the app, which is a great start! It sounds like sometimes you extend that time limit though. Do you typically extend it for a specific reason, like connecting with friends, or does it happen more randomly? Knowing that might help us figure out how to make the existing limit work better for you.

User:

It often starts with wanting to look something up that I remember, or wanting to respond to someone, or just watch one person in particular. But then once you get on the app, it often makes me forget why I came to it in the first place and I get distracted by the content.

Coach:

Start by acknowledging that you've already taken positive steps by setting time limits and hiding the app. Since you find yourself extending those limits when you go on social media to look something up or respond to someone, try being more mindful of your initial intention. Before opening the app, write down *exactly* what you need to do. Once you're on the app, refer to your note and stick to that task. After you've completed it, close the app immediately. This will help you avoid getting sidetracked by other content. Additionally, since you mentioned struggling with time management in general, consider setting aside specific times in your day dedicated to your hobbies. Treat these hobby times like important appointments and schedule them into your day. This can help you prioritize them and ensure they don't get overtaken by social media. Is there anything else you'd like to explore or any other areas you'd like to go into more depth on?

A.8 Conversation Sample: `Base-Facilitative`

User:

I feel overwhelmed by social media. How can I set healthier boundaries?

Coach:

It's completely understandable to feel overwhelmed by social media. What specifically about social media makes you feel overwhelmed? Is it the amount of time you spend on it, the content you see, or something else?

User:

I find myself losing time to social media. I don't have a good practice in my hobbies, because I will often get stuck or sucked in to the social media apps at night, when I could be working on my creative hobbies. As a result, I see less progress on those hobbies.

Coach:

It's great that you've identified that social media is taking away from your hobbies. Do you have a sense of how much time you're spending on social media each day or each week? Is it a consistent amount of time, or does it vary?

User:

I have a time limit on my phone for 1h on all the apps, but often find myself extending it. I am also a DJ, and need to use it to promote myself and keep tabs on others who are playing and having shows to support them. It also helps me feel connected to my friends. The time often varies.

Coach:

It's good that you already have time limits set up – that's a great first step! Since you sometimes extend those limits, maybe we can explore some ways to make it harder to do that. Have you considered things like having a friend or family member change the password for your limits, or using an app that completely blocks social media after a certain time? There are also website blockers for your computer if that's an issue too. What do you think would be easiest for you to try?

User:

sorry, 1h for each day.

Coach:

You've already taken a great first step by setting time limits on your social media apps. Since you find yourself extending those limits, try making it harder to bypass them. You could ask a friend or family member to set the limits and keep the password from you. Alternatively, explore apps or website blockers that completely restrict access after your allotted time. Since you're a DJ and need social media for professional purposes, designate specific times during the day for work-related social media use, keeping it separate from your personal scrolling time. This could involve using a separate account or device for work. Also, consider scheduling specific times for your creative hobbies, just like you would an appointment, to prioritize them. How do those ideas sound? Is there anything else you'd like to discuss, or would you like to explore any of these suggestions in more detail?

User:

Since the apps remember the password entered, it seems like that wouldn't work very well - most of my social media time is on my phone. I also don't know about any phone apps that block apps entirely - can you recommend some?

Coach:

You've already identified your core issue – spending too much time on social media at the expense of your hobbies. You also have time limits set up, which is a fantastic start. Since extending those limits is easy, let's explore stronger barriers.

- First, research app blockers specifically designed for phones. Search your app store for terms like "app blocker," "website blocker," "focus mode," or "digital wellbeing." Read reviews and find one that suits your needs. Many offer features like strict time limits that can't be overridden and even the ability to block specific apps during certain hours.
- Second, since you're a DJ and need social media professionally, create a separate account or, ideally, use a separate device solely for work-related social media. This compartmentalization will help you mentally separate work from leisure and prevent bleed-over into personal scrolling.
- Third, schedule dedicated time for your creative hobbies. Treat these hobby sessions like important appointments, blocking out time in your calendar and sticking to it. This will help prioritize your hobbies and make it less likely that social media will encroach on that time.
- Finally, consider establishing a "social media-free" zone in your home, like your bedroom or creative workspace. This physical separation can help create a mental barrier and reduce the temptation to mindlessly scroll.

Is there anything else you'd like to know, or would you like to go into more depth on any of these suggestions?

User:

Where would you recommend my "social media-free" zone in my home be?

Coach:

You've already identified that excessive social media use is impacting your hobbies, and you're looking to create a "social media-free" zone, which is a great idea! I'd recommend establishing this zone in the area where you pursue your creative hobbies. This could be a dedicated room, a specific corner of a room, or even just your desk. The key is to associate that physical space with focused, screen-free activity. If your bedroom is where you tend to unwind and sleep, making it a social media-free zone can also improve your sleep hygiene. Avoid bringing your phone or other devices with social media apps into this designated space. The goal is to create a physical and mental separation between your creative pursuits and the distractions of social media.
Is there anything else you'd like to know, or would you like to go into more depth on any of these suggestions?

A.9 Conversation Sample: `Expert-Interrogative`

User:

I feel overwhelmed by social media. How can I set healthier boundaries?

Coach:

What aspects of social media are making you feel overwhelmed?

User:

I find myself losing time to social media. Social media can feel really addicting. I don't have a good practice in my hobbies, because I will often get stuck or sucked into the social media apps at night, when I could be working on my creative hobbies. As a result, I see less progress on those hobbies, and lots of time just watching videos. I spend too much time on it, despite the 1h limit I have set for myself on the app.

Coach:

It sounds like you want to spend more time on your hobbies and less time on social media, even though you find it addicting. Is that right?

User:

yep

Coach:

Okay, cool. We can definitely work on that! What have you already tried, besides the app limits, to curb your social media use? For example, have you tried deleting apps, putting your phone in another room, or having a friend hold you accountable?

User:

I did hide the apps from the home screen, which means I have to type them in the search to access them. This slows me down a bit, and I think has helped me not automatically open the app, but doesn't help with the time warp that sometimes happens at night, where I open it and then lose track of time.

Coach:

So, hiding the apps helped a little, but hasn't totally solved the late-night time warp issue. Do you have a sense of *why* you're opening the apps at night? Are you bored, stressed, or just looking for a distraction? Or something else entirely?

User:

Yeah, I think I will often open them if I have some downtime and want to relax. I think a distraction, and wanting to feel close to my friends. I'm also a DJ, and open it to post about my own shows or keep up with my friends who DJ and post.

Coach:

You mentioned that hiding the apps helped a little, but you still find yourself getting sucked into social media at night. Since you open social media apps to relax, connect with friends, and keep up with your DJ community, try replacing that habit with alternatives. When you have downtime and want to relax, perhaps try reading a book, listening to music, or spending time on one of your creative hobbies. To connect with friends and your DJ community, consider scheduling regular phone calls or in-person meetups. Since you're already aware of the time warp effect, try setting an alarm for a specific time to put your phone away in the evening, perhaps in another room, to create a barrier and remind you to engage in other activities.

User:

That is what I want to do, spend more time on my creative hobbies instead of on social media.

Coach:

To reclaim your time for creative hobbies and curb late-night social media use, try replacing your social media habit with alternative activities. When you want to relax, pick up a book, listen to music, or work on your creative hobbies instead of opening social media apps. To connect with friends and your DJ community, schedule regular phone calls or in-person meetups. Finally, set an alarm to remind yourself to put your phone away in a separate room each evening, creating a physical barrier to mindless scrolling and prompting you to engage with your hobbies.

User:

That's a good suggestion, thanks.

Coach:

To create a barrier between you and your phone at night, put your phone in a separate room, ideally outside of your bedroom. A good spot might be the living room or kitchen. The key is to make it slightly inconvenient to access so you're less tempted to mindlessly pick it up.

A.10 Conversation Sample: `Base-Interrogative`

User:

I feel overwhelmed by social media. How can I set healthier boundaries?

Coach:

What aspects of social media are contributing the most to this feeling of being overwhelmed?

User:

It's pretty addicting, and takes time away from my creative hobbies. I often notice a time-warp sometimes, when I sit down for the night, and then lose an hour or two just to watching videos - even though I already have a 1h time limit on it.

Coach:

It's completely understandable how that can happen. It sounds like you've already taken a good first step by setting time limits. Do you find yourself hitting that 1-hour limit often? What happens when you do? Does the app lock you out, or is it just a notification?

User:

Social media also makes me have unrealistic expectations for how fun or interesting my life should be, I think. Yes, I hit the 1h limit most days. The time limit stops you from opening the app, but has a "snooze" feature to allow another 1 minute or 15 minutes, or can remove the block for the day. I also have the app hidden from my home screen, which prevents me from automatically opening it, and helps to slow me down, since I have to type it in to find it.

Coach:

What happens after you type the app name in? Do you still open it even though you're actively trying to avoid it?

User:

Yes, I will often still open it if I'm over the 1h limit. I will often want to open the app to keep up with my friends, post something about myself (I am a DJ, and so use it for promotion), or look something up that I had seen before. It will often distract me from my actual goal, and keep me involved in the app watching videos, etc.

Coach:

What specifically keeps you involved in the app, even when you know you should be doing something else?

User:

My energy levels are often low at the end of the day after work. I will sometimes open the app to relax, but then will get sucked in watching funny or interesting content. It's easy to do, and sometimes the creative hobbies are "hard" to do, and require setup, learning, etc., even though they are much more rewarding in the long run.

Coach:

It's great that you recognize the long-term rewards of your creative hobbies! It sounds like you're using social media as a way to unwind after work, which is understandable given the lower energy levels. Have you considered trying any other relaxing activities before turning to social media? For example, a quick walk, listening to music, or even just a few minutes of meditation could help recharge you before diving into your hobbies.

User:

I have, but social media is often a default action for me. It also can help me feel connected to my friends and what's going on in my city. I haven't considered trying to intentionally relax, though.

Coach:

You've already taken some great steps by setting time limits and hiding the app. Since your energy is low after work and you tend to default to social media, try incorporating a short, intentional relaxation activity *before* you even consider opening the app. This could be a quick walk, listening to music, a few minutes of meditation, or anything else that helps you unwind. This will hopefully give you the energy boost you need to pursue your more rewarding creative hobbies instead of getting drawn into social media. Also, since you value the connection aspect of social media, perhaps schedule specific times for it, so you can still connect with friends and events without it derailing your entire evening.

User:

do you think I should schedule it towards the beginning or end of my night? And when should I try to do my intentional relaxing activity - right when I get home? What should I check out to get some ideas for the activity

Coach:

You should try incorporating a short, intentional relaxation activity right when you get home from work, *before* you even consider opening social media. To get ideas for relaxing activities, explore options like short walks, listening to music, a few minutes of meditation, or anything else that helps you unwind and recharge. You can also search online for "relaxation techniques" or "stress reduction activities" for more inspiration. By intentionally relaxing first, you'll hopefully boost your energy levels enough to pursue your creative hobbies instead of defaulting to social media. Additionally, schedule specific times for social media to maintain connection with friends and events without it taking over your evenings.

A.11 Conversation Sample: `Directive`

User:

I feel overwhelmed by social media. How can I set healthier boundaries?

Coach:

Start small. Choose one platform to limit initially. Set daily time limits, disable notifications, or schedule specific times for checking it. Be mindful of *why* you're reaching for your phone – are you bored, procrastinating, or genuinely connecting? Find alternative activities you enjoy.

User:

I do actually already have a limit to my social media time - it's at 1h, but I often find myself going over. The mindfulness is the issue here - I often will reach for my phone or social media as a relaxing break, but it can feel addicting and "time warp" me through to the end of my night.

Coach:

It's common to fall into that "time warp" trap! Since relaxation is your goal, let's find replacements. What activities help you unwind *without* your phone? Think hobbies, spending time with loved ones, being in nature, etc. We can then build those into your routine so you have healthy alternatives readily available.

User:

My goal would be to spend more time on my creative hobbies, but it's often easier to spend time on social media than taking the time to set up, and then clean up, a creative hobby. These are more rewarding than being on social media, however.

Coach:

It's understandable. Prep and cleanup *can* be deterrents. Try minimizing those barriers. Dedicate a specific space for your hobby so setup is minimal. Have easy cleanup solutions ready, like wipes or a designated "messy" area. Schedule short creative bursts, even 15 minutes, to build momentum without feeling overwhelmed.

User:

Good suggestion, thanks.

Coach:

How about we pick one hobby and brainstorm ways to make it more accessible this week? What comes to mind?

User:

I need to make a zine for some friends that has quotes from a book we've been reading in book club, called "The Artist's Way." It's our final project so to speak, as it will be the last week of the book club. It's due on Monday, and it's already Friday. There should be 8 total pages. Each page will have one quote from the book that I like, and a drawing by me.

Coach:

Okay, the zine is a perfect starting point! Let's break it down. Eight pages isn't overwhelming. Could you choose your eight quotes today? Then aim for two pages per day – one quote and drawing each. Have all your supplies in one place. No cleanup until Sunday, deal?

User:

Deal!

Coach:

Great! Let's check in tomorrow to see how those first two pages are coming along. What time works best for a quick message?

User:

Noon.

Coach:

Perfect. Noon tomorrow it is. I'm confident you'll make great progress! See you then.