# CHAPTER 7 LLMS TO SUPPORT PERSPECTIVE-TAKING AND PROMOTE CHANGE TALK IN VIRTUAL HUMAN CONVERSATIONS

## 7.1 Introduction

### 7.1.1 Overview

In line with the thesis, Study 2 demonstrated that perspective-taking can impact conversational engagement, as reflected in disclosure quantity and depth. Motivated by Study 1, Study 2 also provided evidence that a VH-delivered motivational interview can improve readiness to address mental health. This leaves two primary investigations for the present Study 3. First, there remains an open question regarding how perspective-taking can be more actively supported within VH mental health conversations. The Pre-Study suggested that avatars may not constitute a practical or effective support mechanism in VH conversations. Second, Study 2's demonstrated a significant improvement in readiness through the motivational interview, but it is unclear how perspective-taking and increased engagement contributed to these improvements. It also remains unclear whether perspective-taking can be leveraged to *selectively* elicit, change-aligned engagement that fosters readiness, rather than broader effects on disclosure.

Within the context of perspective-taking interventions, this Study 3 aims to determine (1) how perspective-taking can elicit change-aligned conversational engagement, (2) how LLMs can support perspective-taking, and (3) assess the relationships between perspective-taking, engagement, and readiness. Particularly, Study 3 investigates how increasing manipulations of LLM involvement affect the degree of *overlap* with the taken perspective, the rate of *change talk*, and *readiness* for mental health change. To explore this, I designed a 2x2 study in which students at the University of Florida engaged in a motivational interview that alternated between a *counselor* perspective and their own self-perspective.

### 7.1.2 Research Questions

The following questions are investigated to assess the effect of LLM involvement on perspective-taking overlap, change-aligned conversational engagement, and readiness for mental health change.

- **RQ1$_{\textbf{PerspectiveOverlap}}$** How do increasing manipulations of LLM involvement influence the extent to which users overlap with the taken perspective?
- **RQ2$_{\textbf{ConversationalEngagement}}$** How do increasing manipulations of LLM involvement influence the rate of change-aligned conversational engagements?
- **RQ3$_{\textbf{Readiness}}$** How does the resulting perspective-taking and change-aligned conversational engagement influence users' readiness to change their mental health?

## 7.2 Motivation

This study explores how perspective-taking can elicit greater change-aligned engagement, referred to as change talk. Since motivational interviewing counselors aim to promote change talk in clients (see Section 2.1.2), Study 3 tasks participants with taking the perspective of a counselor and evaluates the effects on their change talk. Specifically, participants engage in a motivational interview in which they self-disclose from their own perspective and, in turn, design a response to their self-disclosure from the perspective of a counselor. This self-reflective conversation design is adapted from prior research exploring motivational interviewing with perspective-taking [11, 10, 9]. In ConVRself, obesity unit participants engaged in a standard motivational interview with a VH or a self-conducted motivational interview, in which they spoke from a self-perspective before responding to themselves from an embodied counselor's perspective [9]. The self-conducted motivational interview was found to be effective in promoting readiness and confidence while reducing uncontrolled or emotional eating behaviors. Because motivational interviewing is designed to help clients hear their own reasons for change, this self-conducted format offers a promising design space for incorporating changes in user perspectives [183]. It also grounds the user in their own perspective throughout the conversation, which was a limitation of the prior Study 2's design (see Section 6.4.4). Accordingly, I build upon their design by incorporating LLMs to help users achieve a counselor perspective and design counselor responses.

Additional context is provided regarding the crafting of the counselor response and the overarching study design and analyses. To help users craft counselor responses, I decompose the process into distinct steps based on key aspects of motivational interviewing. First, motivational

interviewing counselors attend to whether client language reflects openness to change (change talk) or reasons to maintain the status quo (sustain talk), and orient their responses accordingly to promote change and soften resistance. Clients' utterances may also reflect an absence of change or sustain talk; these responses are considered neutral talk. Greater frequencies of clients' spoken change talk have been demonstrated to predict behavioral change (e.g., smoking cessation) [277, 257, 258]. Second, to help clients discuss opportunities to change, motivational interviewing leverages four core strategies (OARS): asking an *open*-ended question, providing an *affirmation*, offering a *reflective* statement, or delivering a *summary* to advance the conversation [276, 279]. Proper usage of OARS strategies has also been shown to predict the presence of change talk in clients' utterances [259]. As a result, this Study 3's motivational interview aims to mimic this process; participants designate a focus and select a desired OARS strategy before writing an utterance to promote change talk.

To evaluate the effects of perspective-taking, change talk, and readiness for change, this study employs a structural path model to quantify their relationships with the included independent variables (IVs). The IVs in this study are based on interaction paradigms described in Section 2.1.3, and broadly refer to who initiates the response (Locus) and whether it is generated independently or collaboratively (Structure) (a more formal description is provided in Section 7.3.1.1). Because path models rely on assumptions of directional influence [286], hypotheses for these variables were developed to build the path model based on prior studies and supporting literature. These hypotheses map directly to the RQs provided in Section 7.1.2. First, I hypothesized that Study 3's IVs for LLM involvement would influence perspective-taking.

**H1** **Locus**, **Structure**, and **Locus** $\times$ **Structure** will directly influence **Overlap**.

I also hypothesized that the IVs and perspective-taking would influence the rate of change talk, based on Study 2's findings.

**H2** **Locus**, **Structure**, **Locus** $\times$ **Structure**, and **Overlap** will directly influence the rate of **Change Talk**.

Finally, I hypothesized that perspective-taking and change talk would influence readiness based on the supporting literature.

**H3** **Overlap** and **Change Talk** will directly influence Δ **Readiness**.

## 7.3 Methods

A 2x2 study design was employed to assess how LLMs can support perspective-taking, change talk, and readiness.

### 7.3.1 Study Design

This section describes the study's IVs, Locus and Structure, and how the IVs represent increasing manipulations of LLM involvement within the intervention.

#### 7.3.1.1 Independent Variables

Two IV dimensions were crossed to form four conditions that varied the extent to which LLMs supported the counselor-response generation process. I employed the human-in-the-loop (HITL) framework to manipulate the balance of user and LLM involvement [284, 281, 113] (see Section 2.1.3 for greater detail on human-AI collaboration). HITL describes the extent of user involvement in automated systems, ranging from fully automated to fully user-driven, aligning with interaction paradigm in which initiative is distributed between human and AI actors [284, 360, 303, 190]. From hereon, the usage of AI in Study 3 is constrained to LLMs. Accordingly, I focus on two IVs: (1) ***Locus*** *of initiative* (**User**-initiated vs. **LLM**-initiated), which indicates whether the user or LLM acts as the initiator in generating counselor-responses, and (2) *initiative* ***Structure*** (**Solo** vs. **Collaborative**), which specifies whether counselor-responses are generated by a single actor or through a mixed-initiative process in which both user and LLM collaboratively contribute [190]. Together, the two IVs model a continuum of automated systems and HITL processes, enabling investigation of how varying degrees of LLM involvement support perspective-taking and conversational engagement. This continuum ranges from fully user-driven interaction, where users independently construct counselor-responses without LLM involvement (**User-Solo**), through user-initiated interaction with LLM guidance (**User-Collaborative**) and LLM-initiated interaction with user guidance (**LLM-Collaborative**), to fully LLM-driven

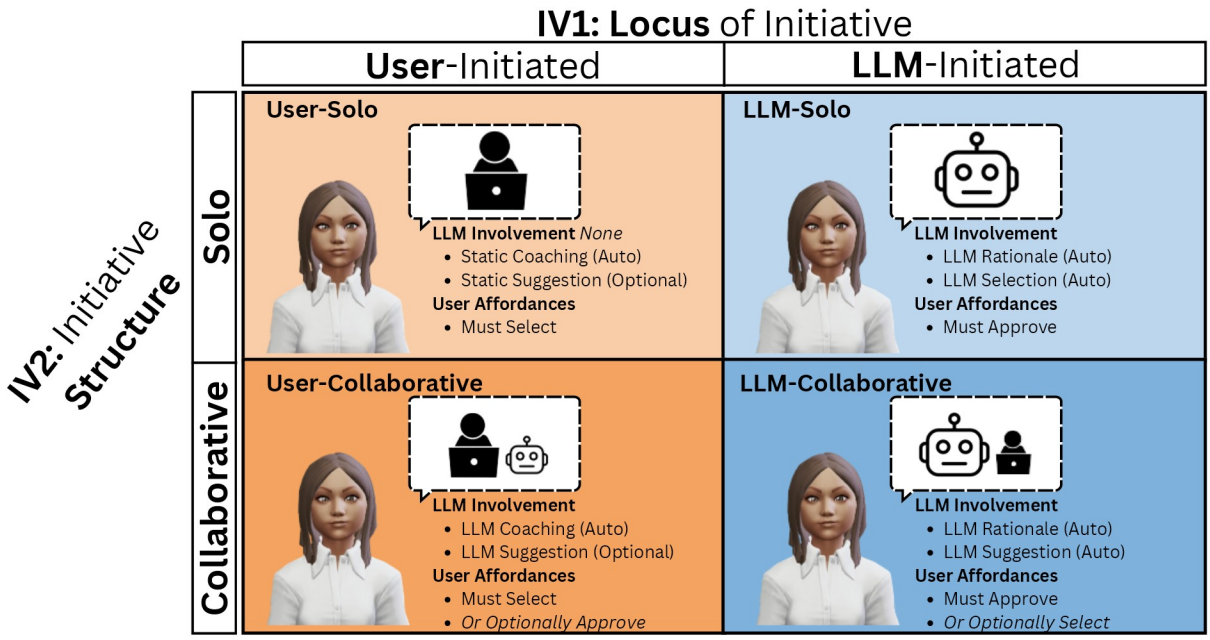


Figure 7-1. Visualization of 2x2 study design with IVs, **Locus** of initiative and initiative **Structure**, their manipulations of LLM involvement, and user affordances with respect to human-in-the-loop.

interaction, where the LLM fully constructs counselor-responses and only notifies the user (**LLM-Solo**) [221]. This design enables evaluation of LLM-scaffolded perspective-taking compared to a non-scaffolded control (User-Solo) (see Figure 7-1 for a high-level overview).

#### 7.3.1.2 Intervention Overview

The motivational interview is modeled as a turn-based interaction with two alternating phases (perspectives): a *self-responding* phase, in which users respond to motivational interviewing prompts from their self-perspective, and a *counselor-responding* phase, in which a response is generated from the counselor's perspective [9]. The IV manipulations occurred during the counselor-responding phases. A broad overview of the counselor-response steps is provided in this section; however, differences in interactions across conditions are further elaborated in Section 7.3.2.2. After submitting their self-response, users were presented with a separate conversation interface and instructed to adopt a counselor's perspective to generate a response intended to facilitate discussion of change. Regardless of condition, generating a counselor-response consisted of four sequential steps: (1) imagining and taking the counselor's perspective, (2)

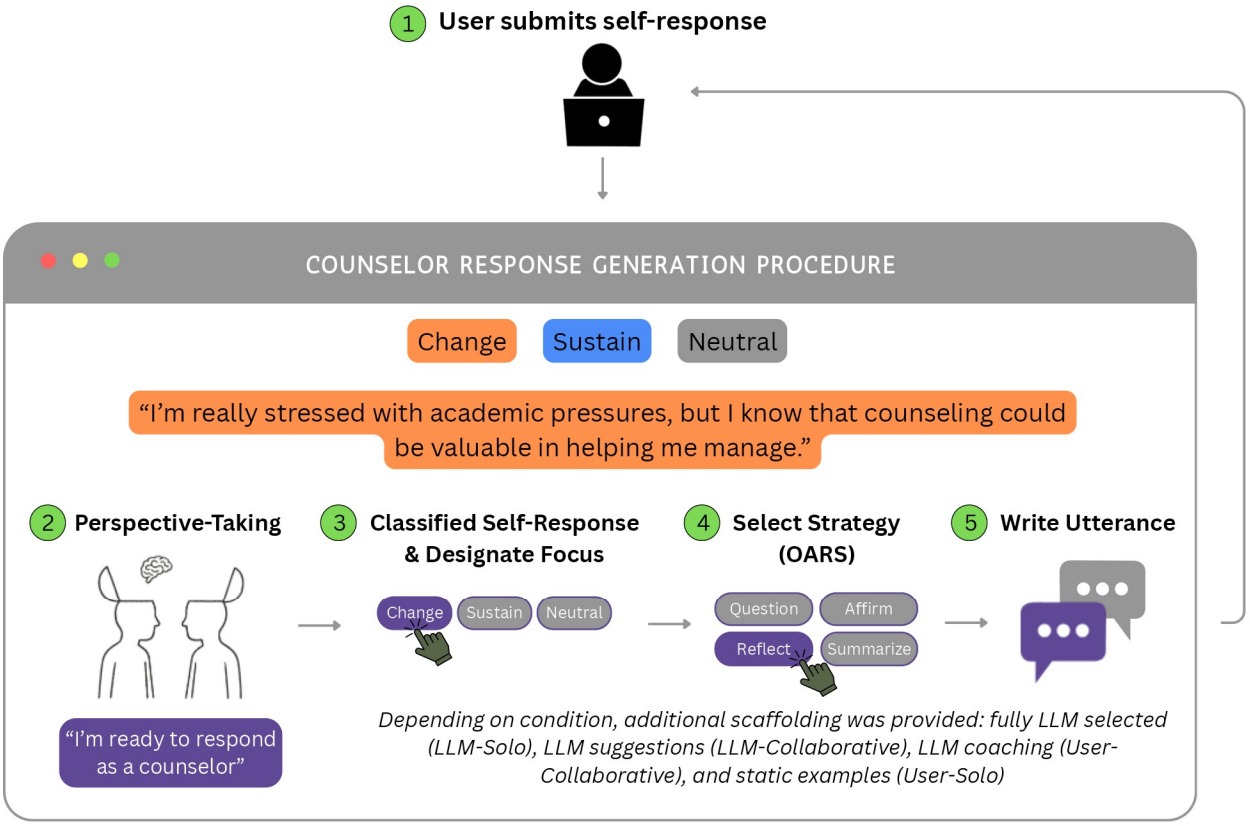


Figure 7-2. A walkthrough on the interaction from (1) user self-response through the counselor-response generation: (2) perspective-taking imaginative task, (3) observing the classified-self response and designating a focus, (4) selecting an OARS strategy, and (5) writing the utterance. Pictured is a self-response classified as *change*, as it demonstrates openness to exploring healthier behaviors.

observing a classified version of their self-response and designating a corresponding focus, (3) selecting an OARS strategy, and (4) writing the utterance (see Figure 7-2).

1. Users were presented with a pop-up modal containing text and visual instructions that prompted a shift to the counselor's **perspective**. Interaction was temporarily paused, and users were instructed to imagine themselves as a counselor responding to their prior self-response to promote change talk.

2. Users were shown a system-classified version of their prior self-response and asked to **designate** a conversational focus. To mirror how counselors attend to clients' speech patterns (Section 7.2), a fine-tuned LLM classifier categorized each self-response as

primarily reflecting change, sustain, or neutral talk (see Section 7.3.2.1). Using this feedback, users designated a conversational focus aimed at amplifying change, softening resistance, or exploring neutral sentiments (likely corresponding to the categorization). This classification was presented as visual feedback to help users select an appropriate focus, while optionally allowing users to redirect the conversation.

3. Afterwards, users selected an OARS **strategy** to guide their response generation. The strategy is oriented around the designated focus (e.g., change-focused open-ended questions may ask how users might take the next steps; whereas sustain-focused open-ended questions may aim to develop the discrepancy and consider where users want to be).

4. Finally, an **utterance** is written for the counselor-response according to the designated focus and selected strategy.

While the process comprises multiple steps, it is streamlined through two multiple-choice selections and a single open-text input.

### 7.3.2 System

To achieve the condition and intervention design described in Section 7.3.1.1, two primary components are described: the talk-type classification model and the overarching motivational interviewing system.

#### 7.3.2.1 Talk-Type Classification

When designating a focus during counselor-response generation, the system applies user-state modeling to infer the motivational orientation expressed in each user utterance, operationalized through a fine-tuned model that classifies speech as change, sustain, or neutral talk. The classified codes are also used in the data analysis of this study. In this section, I describe the creation of this classifier, covering data processing, data preparation, and benchmarking and validation for its use.

**Processing**. To train the classifier, I adopted the *AnnoMI* dataset [406], which contains 133 therapist-annotated motivational interviewing transcripts. Each transcript includes segmented

utterances from both the client and the counselor, along with a labeled talk type (i.e., change talk, sustain talk, or neutral talk) assigned to each client utterance. These annotations follow established motivational interviewing training procedures [285, 276], making the dataset well-suited for modeling approaches [405]. A primary challenge in utilizing *AnnoMI* for this Study 3 is that VH conversations are typically turn-based, whereas *AnnoMI*'s transcripts draw from real-life conversations that are naturally free-flowing and interrupted as clients and counselors build on each other's speech. To address this mismatch, I employed Webb et al.'s process to reconnect temporally divided utterances produced by the same speaker, which has also been shown to improve downstream classification performance [394]. When multiple talk-type annotations were associated with a reconnected utterance, the dominant classification was assigned when a clear majority was present. Importantly, neutral talk does not represent a distinct communicative category but instead reflects the absence of change or sustain talk [285]. As such, reconnected utterances containing both neutral and change or neutral and sustain codes were assigned to the corresponding non-neutral category. Reconnected utterances containing both change and sustain language, however, were not merged; instead, they were preserved as separate conversational turns to avoid conflating motivational signals. Although this preprocessing reduces the quantity of data points, it preserves the aims of the annotating process. Furthermore, since the dataset is heavily skewed toward neutral talk, this approach ensures that comparatively sparse but theoretically critical change and sustain utterances are retained and appropriately represented. The processed dataset included (n = 3937) codes, of which 26.7% were change talk, 12.5% were sustain talk, and 60.8% were neutral talk.

**Preparation**. Although the *AnnoMI* dataset provides a strong foundation, the available training data remain limited, especially given the reduced number of utterances resulting from the turn-reconstruction process. To address this limitation and support model performance, I employed back-translation techniques to synthetically generate additional training data from the processed dataset [21]. Back-translation entails translating monolingual text into a secondary language and subsequently translating it back into the original language. This process introduces

controlled lexical and syntactic variation while preserving the underlying semantic intent of the utterance [351]. I used the Marian Machine Translation (MarianMT) models for English-to-French and French-to-English back-translation [204, 48]. The availability of large, high-quality parallel corpora and the relative maturity of translation models for English and French motivate their use in this study [57, 367]. This process doubled the total number of included data points to 7874 codes. After stratified shuffling at the transcript level to preserve class balance, the dataset was split into 80% training/validation and 20% test sets. All conversation turns within a transcript were kept in their original order, and back-translated transcripts were assigned to the same split as the original to mitigate overfitting.

**Benchmarking and Validation**. Similar to prior work using classifiers in motivational interviewing agents [291], I provide brief benchmarks of three models' performance on the resulting dataset. These models include the base Bidirectional Encoder Representations from Transformers (BERT Base) model, a layer-frozen BERT model updated using lightweight adapter modules (BERT Adapters), and the latest fine-tunable model from OpenAI (GPT-4.1 Mini). The BERT models are selected based on previous findings on the *AnnoMI* dataset, which reported the greatest performance relative to baseline approaches (e.g., Random Forest, Convolutional Neural Networks) [405]. To evaluate model performance, F1-scores are reported both as a macro-averaged measure and on a per-class basis. The F1-score is a commonly used evaluation metric in machine learning, defined as the harmonic mean of precision and recall, and is employed to provide a balanced assessment of performance when both false positive and false negative errors are important (like in classification) [80, 262]. Optimal hyperparameters were found using integrated tools or libraries (e.g., Optuna) before evaluation. The BERT models reported highly similar scores to the previous models generated with *AnnoMI* [405]. Based on the brief benchmarking, I adopted the best-performing model for this study: the fine-tuned GPT-4.1 Mini model (F1-Macro = 0.68) (see Table 7-1).

While the fine-tuned model outperforms previous models [405], its use in this study's interactions and data processing is motivated by factors beyond F1-score performance. First, the

Table 7-1. Table depicts the F1-scores for each model overall (macro) and per class (neutral, change, sustain). The best performing model was the fine-tuned, prompt-engineered GPT-4.1. Accuracy is also included to help contextualize results.

| | Talk Type Prediction | | | | |
|---|---|---|---|---|---|
| **Model** | **F1-Macro** | **F1-Neutral** | **F1-Change** | **F1-Sustain** | **Accuracy** |
| GPT-4.1 (Fine-Tuned) | 0.68 | 0.81 | 0.64 | 0.60 | 0.74 |
| GPT-4.1 (Prompt-Only) | 0.60 | 0.78 | 0.49 | 0.54 | 0.68 |
| BERT (Adapters) | 0.55 | 0.73 | 0.47 | 0.44 | 0.62 |
| BERT (Base) | 0.55 | 0.76 | 0.51 | 0.38 | 0.65 |

models' performance is consistent with the inherent difficulty of annotating spoken motivational interviews, where class boundaries are often ambiguous and highly context-dependent. Even among therapists, only moderate agreement was observed across client talk type codes in *AnnoMI* (Fleiss' $\kappa = 0.47$) [406]. Second, there is a reasonable expectation for higher model performance in this Study 3's LLM-supported conversations. Related to the first factor, the *AnnoMI* authors noted that the dataset contains a large proportion of short, context-dependent utterances (e.g., therapist-spoken acknowledgments such as "mm-hmm" or "right") that limited agreement and constrained turn-level classification performance. Such responses can be typical in spoken human-human conversations, whereas prior work has shown that conversational dynamics differ in typed human-AI conversations [185], with AI responses tending to be substantially longer than those of healthcare professionals [18]. I also conducted brief ablation testing on the dataset, which confirmed higher performance with lengthier therapist responses. Finally, rather than relying on raw utterance-level classifications for data analysis, I operationalize change talk as an aggregate *conversation-level* measure. Classification uncertainty primarily reflects boundary ambiguity (i.e., change-neutral or sustain-neutral) rather than systematic polarity reversal, making conversation-level aggregation more appropriate than interpreting individual utterances (see Appendix Figure C-1). Further details and statistical validations of the coded talk-types are provided in Section 7.3.3.2.

#### 7.3.2.2 Motivational Interview

This Study 3 intervention utilizes the underlying VH authoring platform architecture described in Chapter 5 (BABY). Within this architecture, this section provides a brief overview of

the counselor-response generation, available-on-request resources, and the implementation differences of the IVs.

**Overview of Counselor-Response Generation**. To provide parity across conditions, a fully automated pipeline was designed and serves as the foundational process, from which condition-specific manipulations were derived. The fully automated pipeline is the process used in LLM-Solo and is described here.

After generating the classification for the submitted self-response, the system identifies a conversational focus, selects an OARS strategy, and generates the counselor's utterance. Under full automation, focus designation requires no additional inference, as it is directly derived from the classified user utterance. To determine which strategy to employ, the system provides the designated focus and the user's prior self-response to a reasoning model equipped with tool-calling, which is prompted to select the response strategy most likely to encourage subsequent change-oriented language. Reasoning models use chain-of-thought reasoning to decompose complex reasoning tasks into intermediate steps that can be iteratively assessed to improve response quality [395]. Prior work has shown that delegating such intermediate steps to specialized *tools* can further enhance reasoning performance and controllability in complex question-answering tasks, computations, or even language translation [302, 341]. In the context of motivational interviews, counselors must navigate often-sensitive conversations by strategically picking OARS strategies to cultivate change talk and soften sustain talk [87]. Therefore, this study implements 12 "tools" (a change, sustain, and neutral variant of each OARS strategy), which are selected via the reasoning model (GPT-5-mini). The designated conversational focus constrains this selection to the corresponding subset of OARS strategies, and the model is prompted to use reasoning in selecting the "tool" (strategy) most likely to promote change talk in the subsequent turn, given the OARS strategy definition, examples, and conversation history. As in Study 2, this Study 3's motivational interview strategies, definitions, and examples are taken verbatim from several sources, including motivational interviewing books [275, 279], taxonomies or findings on motivational interviewing strategies [175, 324], and educational materials for motivational

interviews [230]. Broadly, these entail the following: (1) change-focused strategies to explore approaches to change, reinforce current behaviors, or identify behaviors to adopt/exclude, (2) sustain-focused strategies to emphasize self-autonomy, reframe reluctance with a positive twist, or establishing importance/values to change, and (3) neutral-focused strategies to explore other conversation topics for change, normalizing ambivalence, summarizing user perceptions of the problem, or acknowledging both sides to change.

Alongside the selected strategy, the LLM provides a chain-of-thought justification for it, which is then reformulated into a concise, user-facing explanation via a secondary prompt. This design enables free-flowing conversation that builds on past motivational interviewing systems that rely on preset topics or scripts [59, 291], while also adding dimensions of generative reasoning beyond keyword-triggered strategy selection [306]. With the designated focus and selected strategy, a few-shot example prompt is used to generate the counselor-response utterance. This prompt uses the same definitions, examples, and context as in the strategy selection. In addition to the supporting literature, the interaction was developed based on standard motivational interviewing training, pilot tested, and reviewed by external motivational interviewing experts to provide evidence of face validity.

**Resource Requests**. Beyond *typing* self-responses and generating counselor-responses, users could request resources through *button input* during the self-response phase. This feature was included in response to feedback from participants in Study 2, who noted that while the intervention supported reflection, it lacked clear pathways to concrete support. Given prior literature demonstrating that AI systems may hallucinate or provide unvetted recommendations [199, 132, 296], the system integrated a curated set of six institutionally provided or endorsed resources at the University of Florida, spanning both counseling services (case management, group counseling, short-term counseling) and alternative self-guided supports (campus events, university-endorsed apps, basic needs assistance). When resources were requested, the same tool-calling and reasoning model was used to identify an actionable resource intended to support consideration of change, based on the ongoing conversation and expressed needs. Users could

request a maximum of three resources, which could only be requested via a designated interface button (see Figure 7-3 (Top-Left)).

**Independent Variable Manipulations**. This section describes the 2x2 IV manipulations across the motivational interviewing system, with a focus on the HITL framework. Before crafting the counselor-response, all users re-observed their submitted self-response, now color-coded according to the model's classification output (see Figure 7-3 for interfaces). The manipulations for crafting the counselor-response with LLM involvement are described below (see Figure 7-1 for summary).

- **LLM-Solo** *"Fully LLM"*. In this LLM-initiated, Solo-structure condition, the LLM provides scaffolding by fully executing the motivational interviewing pipeline described in the prior section. The system is responsible for designating the focus using the classifier, selecting an OARS strategy via tool-calling, and generating the corresponding utterance using a few-shot prompt. At each step, a user-facing rationale is provided; however, users must *approve* the resulting response, reflecting minimal HITL involvement and a fully LLM-driven condition [221].

- **LLM-Collaborative** *"LLM, with a touch of user"*. In this LLM-initiated, Collaborative-structure condition, the LLM executes the same counselor-response pipeline as in LLM-Solo but presents each step as a modifiable suggestion with a user-facing rationale. The LLM is initially responsible for selecting the focus, strategy, and utterance, while the user provides oversight to support their decision-making. Consistent with established HITL oversight affordances [113], users must *approve* each step's suggestion, or *optionally self-select* by manually editing the suggestion or requesting a re-suggestion.

- **User-Collaborative** *"User, with a touch of LLM"*. In this User-initiated, Collaborative-structure condition, the LLM executes the same counselor-response framework as in LLM-Solo but presents each step as a user-facing set of coached instructions. Users are initially responsible for selecting the focus, strategy, and utterance,

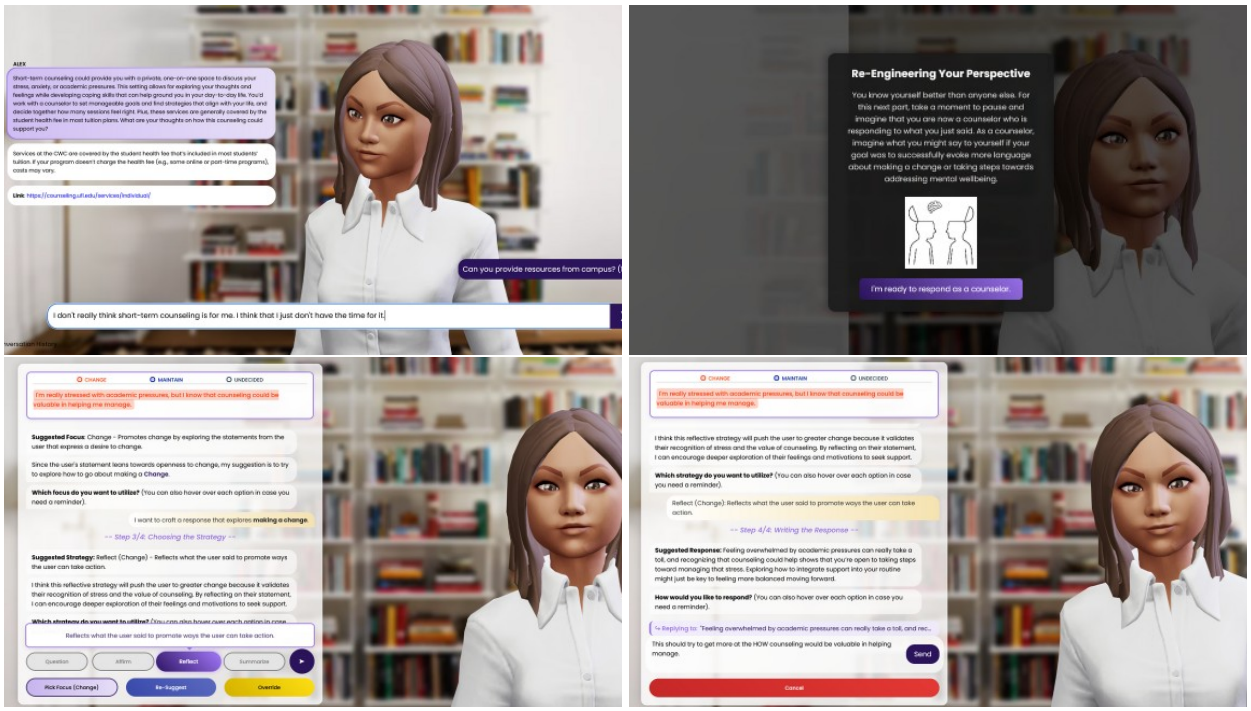


Figure 7-3. Sample interface images of the motivational interview. Pictured is the user submitting a self-response (Top-Left) and the perspective-taking reminder modal (Top-Right). Additionally, the **LLM-Collaborative** interface is shown for selecting a counselor-response strategy (Bottom-Left) and re-suggesting a counselor-response utterance with feedback (Bottom-Right).

while the LLM provides information and support to guide their selection. To achieve this, the LLM-Solo pipeline is employed to internally generate the fully automated selection and rationale. LLM prompting converts the selection and rationale into procedural and metacognitive coaching to reduce cognitive burden compared while preserving user autonomy [151]. Users must manually *self-select* each step using the coached instructions, or can request the suggestion that the coaching was based on and *optionally approve* it, reflecting a higher degree of HITL involvement [221].

- **User-Solo** *"Fully user"*. In this User-initiated, Solo-structure condition, the LLM is not involved. Instead, the LLM is fully removed from the counselor-response framework, and all response selection is performed by the user. Users are responsible for *selecting* the focus, strategy, and utterance, supported only by static coaching and example content provided for

each step. The resulting counselor response is entirely user-authored, illustrating perspective-taking in the absence of LLMs and representing a fully human-driven baseline condition.

### 7.3.3 Measures

To address the **RQs** in this study, analyses were conducted in three stages. First, manipulation checks were performed to verify that the experimental manipulations functioned as intended. Second, the primary analyses consisted of a structural path model examining the relationships among the independent variables, perspective-taking overlap, change talk, and pre-post changes in readiness. Finally, other measures related to the system and attitudes were analyzed to contextualize the primary findings.

#### 7.3.3.1 Manipulation Checks

**Perceived Proactivity and Reactivity**. Manipulation checks were administered to confirm that system behaviors were perceived as intended across conditions. Since the IV manipulations are directly related to automation and initiative, two three-item questionnaires measuring perceived system *proactivity* and *reactivity* were used [107, 174]. Conditions with an LLM-initiated Locus were expected to be perceived as more proactive, whereas conditions with a Collaborative-initiative Structure were expected to be perceived as more reactive.

#### 7.3.3.2 Path Model of Overlap, Change Talk, and Readiness

**Overlap**. While Study 2 indirectly assessed perspective-taking through changes in behavioral engagement, the present Study 3 directly measures perspective-taking using the *Inclusion of Other in the Self* (IOS) scale [13]. Similar to the Pre-Study embodiment questionnaire assessing avatar self-identification, the IOS scale measures the closeness, or *overlap*, with another perspective [139] through an 11-point, pictorial scale [19] (see Appendix C-2). This validated scale has been employed in psychology to assess interpersonal relationships [13], as well as in HCI research to measure self–other overlap when establishing empathy in VR contexts [5].

**Change Talk**. *Change talk* is operationalized as the rate (mean) of change-aligned self-response utterances across the conversation. Given moderate utterance-level uncertainty in

the classifier and variance in conversation turns, raw utterance-level counts of change talk may be unstable and confounded by interaction length. Instead, conversation-level, aggregate estimates provide a more conservative, but appropriate basis for comparison across participants. The change talk codes (along with sustain talk and neutral talk) are captured through the fine-tuned classification model described in Section 7.3.2.1.

To further validate these codes for analysis, I employed "no gold standard" approaches to compare model and human outputs in settings where ground-truth labels are unavailable or inherently ambiguous, consistent with prior models used in motivational interviewing and broader health [373, 273, 368]. I conducted a post hoc analysis of the conversation transcripts collected during the study and coded users' responses (alongside another HCI researcher) into the three client talk types. Although prior work has demonstrated that motivational interviewing annotations can be reliably produced by non-clinician annotators [312], we first established agreement with the therapist-coded annotations in *AnnoMI* using the Motivational Interviewing Skills Code framework [276]. Afterwards, each coder independently annotated a randomized, stratified 20% (n = 255) subset of conversation turns from this study. The process yielded significant agreement between the human and model codes (Fleiss' $\kappa = 0.71$), which reflects *substantial* agreement [228]. Accordingly, the model-derived codes are used as analytical representations of change talk in this study, with the understanding that these codes reflect automated estimates given the potentially ambiguous nature of motivational interviewing annotations.

$\Delta$ **Readiness**. The prior readiness measure used in Study 2 is included in the present Study 3 (see Section 6.3.3.2). Readiness for mental health change is assessed through the *Readiness-to-Change Questionnaire* adapted for mental health [181]. A composite readiness score is captured in both the pre-survey and post-survey of this study, which is used to calculate $\Delta$ Readiness (post $-$ pre). The $\Delta$ Readiness captures differences in readiness for mental health change between pre- and post-intervention assessments.

#### 7.3.3.3 System and Attitudes

**Time**. *Time* is measured as the duration from when a participant accessed the web-based intervention to when they submitted a goal to end the conversation. Because certain conditions involve greater LLM involvement, time is included to contextualize path model findings.

**Conversation Turns**. A *conversation turn* was defined as a single instance in which a participant's typed self-response was paired with a corresponding counselor response. These together are summed for each participant to reflect their conversation turns.

**Requested Resources**. As users could optionally request resources via a button input, *requested resources* is measured by the number of instances in which the request button is used (max = 3). It should be noted that button inputs for resources were assigned a default change-talk code but were not included in analyses of change talk or conversation turns, as there was no user-typed input.

**Attitudes towards the Psychological Online Intervention**. The APOI Questionnaire utilized in Study 2 is employed in this Study 3 [363] to measure the four constructs: *skepticism* and perception of risks, *confidence* in effectiveness, *technologization* threat, and *anonymity* benefits (see Section 6.3.3.3).

**System Usability**. Because IV manipulations may naturally reduce user autonomy or require greater user effort, *system usability* is measured to assess users' perceptions of ease of use and overall interaction quality. System usability is measured using the System Usability Scale (SUS), the most widely used and validated scale for this construct [237, 27, 58]. The SUS contains 10 items (5-point scale) that are combined to form a raw usability score ranging from 0 to 100.

### 7.3.4 Procedure

This section describes the study procedure, from recruitment and pre-survey through the tutorial and motivational interview, and concluding with the post-survey. I conducted a between-participant study using the described system with computer science course-enrolled students at the University of Florida. Participants were recruited through the university's SONA recruitment platform, as in the prior studies. After providing informed consent, participants

completed the pre-survey containing the pre-measures (readiness, intentions). Participants were randomly assigned to one of four conditions (LLM-Solo, LLM-Collaborative, User-Collaborative, User-Solo) and redirected to the intervention, which was hosted as a web-based application built in Node.js using the infrastructure described in Section 5.3.2. Prior to the intervention's start, participants selected a male or female VH for the remainder of the interaction (the same models from Study 2). The VH introduced itself and provided an overview of the study based on the participant's assigned condition. The intervention was described to participants as a motivational interview in which they would explore opportunities for mental health change, including counseling, by learning about conversational techniques used by counselors and designing subsequent counselor-responses.

The motivational interview began with two preset questions and an interactive tutorial to teach users how to use the system. The VH delivered two preset questions that established the conversational focus on one current wellbeing concern identified by the participant and their prior experiences addressing it on their own, similar to Study 2. Afterwards, participants engaged in a guided, interactive tutorial that walked participants through each step of crafting a counselor-response, which could be completed multiple times. The tutorial consisted of an intentionally exhaustive walkthrough of the participant's assigned condition, designed to ensure a clear understanding of the interface capabilities and affordances at each step of the interaction. Specifically, the tutorial provided instructions as users completed each step to create a counselor-response with a sample response: (1) adopting the counselor's perspective; (2) observing sample classified responses and distinguishing between *change*, *sustain*, or *neutral* labels; (3) designating a conversational focus for the counselor response; (4) learning the four OARS strategies and locating tooltips associated with each; (5) selecting an OARS strategy; (6) composing a response aligned with the selected focus and strategy, and (7) returning to their self-perspective. Within each crafting step, the respective condition affordances (e.g., LLM-Collaborative: Approve, Re-Suggest, and Edit) were also demonstrated.

After completing the tutorial, participants initiated the intervention proper by crafting a counselor-response to their submitted self-response for the second preset question. A reduced version of the tutorial was provided for this initial turn to support the transition from the comprehensive tutorial to independent interaction. Following the initial counselor response, participants were required to complete a *minimum of five* additional typed conversation turns, excluding resource-request turns. At any point following the minimum required turns, participants could end the conversation by identifying a next-step goal, consistent with the client-driven nature of motivational interviewing.

Participants were redirected to a post-survey following the motivational interview. The post-survey collected the perspective-taking overlap measure, manipulation checks, and the remaining outcome variables for readiness, intentions, attitudes, and system usability. Finally, participant demographic information was captured in the post-survey, and credit compensation was provided after completion.

### 7.3.5 Participants

An a priori power analysis was conducted using G*Power 3.1 for a 2x2 between-participants ANOVA (fixed effects; main effects and interaction). Assuming $\alpha = 0.05$, power $(1 - \beta) = 0.95$, and a medium effect size of $f = 0.25$, the analysis indicated a required total sample size of $N = 210$ ($n = 53$ per condition) to detect a main or interaction effect. This study was approved by the University of Florida Institutional Review Board, and all participants provided written informed consent before completing the study. To account for dropout and errors in completion, a total of 261 participants were recruited via the research recruitment platform and completed the entirety of the procedure in Section 7.3.4. Seventeen "bad actors" (n = 17) were manually excluded from analyses. Exclusion criteria included repeated nonsensical self-responses (n = 15; e.g., random character strings) or *blatant* disregard for task instructions (n = 2; e.g., one user repeatedly threatened the VH with violence). The final analysis included 244 participants: LLM-Solo (n = 64), LLM-Collaborative (n = 63), User-Collaborative (n = 56), and User-Solo (n =

61). Participant ages ranged from 18 years to 56 years ($M = 22.05, SD = 4.57$); further demographic details, self-reports on mental health, and selected VH are provided in Table 7-2.

Table 7-2. Breakdown of (n = 244) participants by gender, race, and major, as well as reports of counseling received over the prior 12 months, need for support over the prior 12 months, and selected VH.

| Demographics | *N* | % |
|---|---|---|
| **Gender** | | |
| Male | 160 | 65.6 |
| Female | 78 | 32.0 |
| Other | 6 | 2.4 |
| **Race** | | |
| White | 127 | 52.1 |
| Asian | 78 | 32.0 |
| Black | 10 | 4.1 |
| Mixed | 14 | 5.7 |
| Other | 15 | 6.1 |
| **Major (Education)** | | |
| Computer Science | 205 | 84.0 |
| Computer Engineering | 33 | 13.5 |
| Other STEM | 6 | 2.5 |
| **Received Counseling (Past Year)** | | |
| No | 176 | 72.1 |
| Yes | 64 | 26.2 |
| Not reported | 4 | 1.7 |
| **Perceived Need for Support (Past Year)** | | |
| Strongly agree | 30 | 12.3 |
| Agree | 32 | 13.1 |
| Somewhat agree | 50 | 20.5 |
| Neither agree nor disagree | 32 | 13.1 |
| Somewhat disagree | 20 | 8.2 |
| Disagree | 49 | 20.1 |
| Strongly disagree | 31 | 12.7 |
| **Virtual Human Gender** | | |
| Female | 136 | 55.7 |
| Male | 108 | 44.3 |

## 7.4 Results

Analyses proceeded in three stages: (1) group comparisons on manipulation checks; (2) path modeling to evaluate the hypothesized relationship and subsequent comparisons; and (3) group comparisons on additional system and attitudinal measures. Data preprocessing was

conducted in Python (3.12.2), and statistical analyses were primarily performed in R (4.5.0). ART ANOVAs and path modeling were conducted in R using ARTool [401] and lavaan package [334], respectively. For all group comparisons, assumptions for two-way ANOVAs with factors Locus and Structure were evaluated prior to analysis. Shapiro-Wilk tests revealed violations of normality for the measures, $p < 0.05$. Therefore, aligned rank transform (ART) ANOVAs, with Locus (User-initiated vs. LLM-initiated) and Structure (Solo vs. Collaborative) as factors, were conducted for each measure. Effect sizes were calculated and listed via partial eta squared $\eta_p^2$ and rank-biserial correlation ($r$) for the ART ANOVA and post hoc tests. Post hoc procedures followed standard direct-contrast and model-based practices. Consistent with Study 2, ART-C was used for post hoc direct-condition contrasts on manipulation checks and system and attitudinal measures [401, 119]. Because path models estimate effects jointly within a dependency structure, post hoc analyses for path modeling instead require comparisons of estimated marginal means (EMMs) [236]. All post hoc comparisons were adjusted using the Holm procedure to control the familywise error rate [189].

Path modeling was performed by constructing the described initial model and sequentially trimming it. Following the procedures for structural path models [208] and past HCI research [255, 268, 388], hypotheses were tested by first specifying an initial path model that extended beyond the hypothesized relationships, including all possible direct paths among the IVs, overlap, change talk, and readiness. Stepwise model trimming procedures were applied based on theoretical rationale and empirical indicators, including path significance and model parsimony. Models were estimated using robust maximum likelihood (MLR) to accommodate non-normality. Model fit was evaluated using multiple indices, including the Comparative Fit Index (CFI), Tucker–Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA), Standardized Root Mean Square Residual (SRMR), and the robust Chi-square ($\chi^2$) test of exact fit [208]. Good to excellent fit is indicated by CFI and TLI values $\geq 0.95$, RMSEA values $\leq 0.05$, and SRMR values $\leq 0.08$. Additionally, non-significant $\chi^2$ ($p > 0.05$) indicates acceptable model fit since the model-implied covariance matrix does not significantly differ from the observed data [218]. After

producing the final parsimonious model, indirect effects were estimated using nonparametric bootstrapped confidence intervals with 5,000 resamples [179, 255]. Because indirect effects are products of multiple path coefficients and are not normally distributed, statistical significance was determined by whether the 95% bootstrapped *confidence interval* excluded zero [179].

### 7.4.1 Manipulation Checks

**Perceived Proactivity**. ART ANOVA revealed a significant main effect of Locus, F(1, 238) = 5.69, $p = 0.018$, $\eta_p^2 = 0.023$. Regardless of Structure, post hoc analyses revealed that LLM-initiated Loci (Mdn = 4.00, M = 3.56, SD = 0.781) reported significantly *greater* perceived proactivity than User-initiated Loci (Mdn = 3.33, M = 3.37, SD = 0.851), $t(238) = 2.39$, $p = 0.018$, $r = 0.153$. No other significant effects were found.

**Perceived Reactivity**. ART ANOVA revealed a significant main effect of Structure, F(1, 238) = 7.10, $p < 0.01$, $\eta_p^2 = 0.029$. Regardless of Locus, post hoc analyses revealed that Collaborative Structures (Mdn = 3.67, M = 3.63, SD = 0.760) reported significantly *greater* perceived system reactivity compared to Solo Structures (Mdn = 3.33, M = 3.33, SD = 0.813), $t(238) = 2.66$, $p < 0.01$, $r = 0.170$. No other significant effects were found.

### 7.4.2 Path Model

#### 7.4.2.1 Overlap, Change Talk, and Readiness

I developed an initial saturated model comprising the observed exogenous variables (Locus, Structure, Locus × Structure) and endogenous variables (perspective overlap, rate of change talk, and Δ Readiness). The saturated model contained zero degrees of freedom and included all theoretically plausible paths from exogenous to endogenous variables, with perspective-taking overlap and change talk specified as sequential mediators of the effects of Locus and Structure on readiness. Naturally, the initial saturated model demonstrated excellent fit (CFI = 1.00, TLI = 1.00, RMSEA = 0.00, and SRMR = 0.00) (Figure 7-4 contains a reference image of the initial saturated model alongside the final reported path model), and was therefore refined to create the most parsimonious, or minimal, yet theoretically-justified path model. To achieve this, the following steps were systematically applied to trim paths between variables.

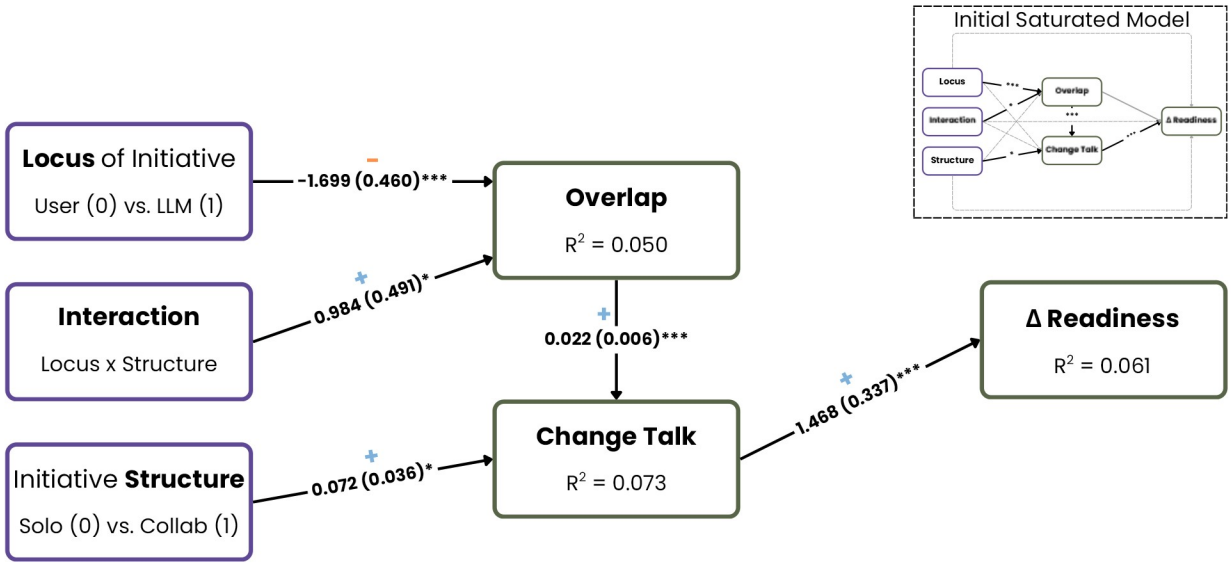


Figure 7-4. The trimmed path model that demonstrates how levels of Locus and Structure (and their Interaction) influence overlap, change talk, and changes in readiness. Only significant paths remain in the trimmed model (*** < 0.001, * < 0.05), and paths illustrate both regression coefficient and (standard error). For IVs, coefficients and standard error reflect (1) relative to (0), where positive values indicate higher outcomes for (1). $R^2$ values are displayed below each endogenous variable name. The initial saturated model is shown in the top-right corner for reference.

1. First, direct paths to the final outcome, Δ Readiness, were examined and demonstrated that only Change Talk emerged as a significant predictor. Since Change Talk significantly influenced Δ Readiness, direct paths from the three exogenous (Locus, Structure, and Locus × Structure) and Overlap paths to Δ Readiness were trimmed, as were non-significant and theoretically mediated through Change Talk.

2. Predictors on the mediator, Change Talk, were then examined. Since Overlap and Structure significantly influenced Change Talk, direct paths from exogenous variables of Locus and the Locus × Structure Interaction were trimmed, as they were non-significant and theoretically mediated through Overlap.

3. Within the final step, the path from Structure to the endogenous variable, Overlap, did not reach statistical significance and was subsequently trimmed for parsimony. The resulting model contained only significant paths and is shown in Figure 7-4. At each stage of

Table 7-3. Table illustrating the direct relationship paths from the trimmed path model, as well as the identified indirect effects. Alongside unstandardized coefficients (B) and standardized coefficients $\beta$, standard error (SE) and significance are shown. Indirect effects are evaluated through 95% confidence intervals rather than p-values (all listed paths are significant).

| Path | B | $\beta$ | SE | Significance |
|---|---|---|---|---|
| **Overlap** ($R^2 = 0.050$) | | | | |
| Locus→Overlap | -1.699 | -0.271 | 0.460 | $p < 0.001$ |
| Interaction→Overlap | 0.984 | 0.138 | 0.491 | $p = 0.045$ |
| **Change Talk** ($R^2 = 0.073$) | | | | |
| Structure→Change Talk | 0.072 | 0.124 | 0.036 | $p = 0.045$ |
| Overlap→Change Talk | 0.022 | 0.232 | 0.006 | $p < 0.001$ |
| Indirect Effects | | | | |
| Locus→Overlap→Change Talk | -0.037 | -0.063 | 0.014 | [-0.072, -0.014] |
| Interaction→Overlap→Change Talk | 0.021 | 0.094 | 0.013 | [0.002, 0.053] |
| **Readiness** ($R^2 = 0.061$) | | | | |
| Change Talk→Readiness | 1.488 | 0.247 | 0.337 | $p < 0.001$ |
| Indirect Effects | | | | |
| Locus→Overlap→Change Talk→Readiness | -0.054 | -0.016 | 0.026 | [-0.126, -0.018] |
| Interaction→Overlap→Change Talk→Readiness | 0.031 | 0.008 | 0.021 | [0.004, 0.089] |
| Structure→Change Talk→Readiness | 0.106 | 0.031 | 0.062 | [0.007, 0.253] |
| Overlap→Change Talk→Readiness | 0.032 | 0.057 | 0.012 | [0.013, 0.061] |

development, the model exhibited a non-significant chi-square statistic; specifically, the final model was consistent with the observed data $\chi^2(7) = 6.59$, p = 0.473. This indicates that the hypothesized model does not significantly deviate from the observed covariance matrix, representing an *excellent* absolute fit: CFI = 1.000; TLI = 1.019; RMSEA = 0.000; and SRMR = 0.027. With respect to the hypotheses used to create the path model, the following direct effects were supported:

**H1** **Locus**, ~~Structure,~~ and **Locus** × **Structure** directly influence **Overlap**.

**H2** ~~Locus,~~ **Structure**, ~~Locus × Structure,~~ and **Overlap** directly influence the rate of **Change Talk**.

**H3** ~~Overlap and~~ **Change Talk** directly influences Δ **Readiness**.

Following identification of the final trimmed path model, indirect effects were examined to assess whether overlap and change talk mediated the relationships between the IVs and Δ

Readiness. Table 7-3 illustrates the significant effects from the created path model, as well as the resulting significant indirect effects. All direct and indirect paths within the trimmed path model were statistically significant, demonstrating that IVs and overlap exert significant indirect effects on Δ Readiness through change talk.

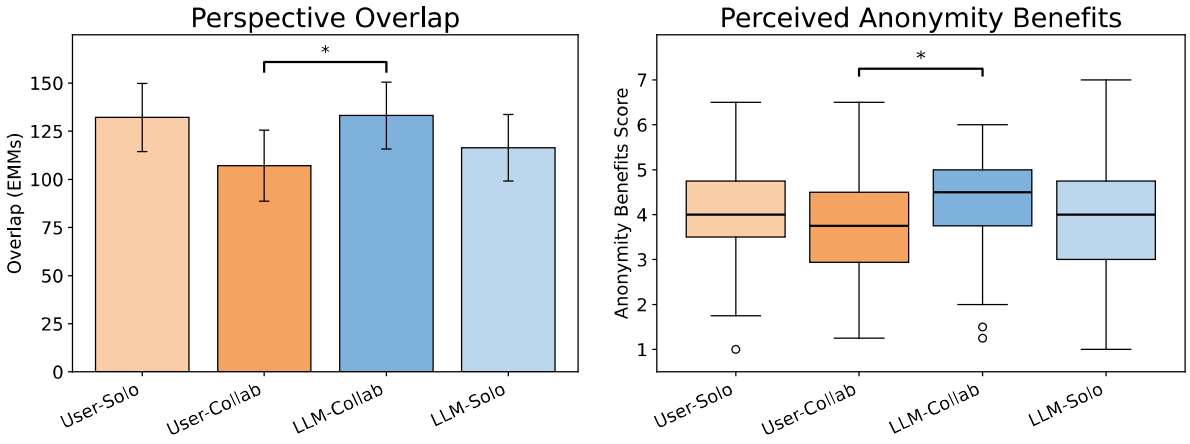


Figure 7-5. Comparisons of Perspective Overlap (EMMs) and Perceived Anonymity Benefits across conditions. A significant Locus × Structure interaction was found in both, demonstrating greater Perspective Overlap and Perceived Anonymity Benefits for LLM-Collaborative compared to User-Collaborative (* < 0.05).

#### 7.4.2.2 Analyses on Overlap

To further evaluate the identified direct interaction effect on overlap, an ART ANOVA was conducted. ART ANOVA revealed a significant main effect of Locus on overlap, F(1, 240) = 8.33, $p < 0.01$, $\eta_p^2 = 0.034$. Regardless of Structure, post hoc analyses revealed that LLM-initiated Loci (Mdn = 5.00, M = 5.47, SD = 2.82) reported significantly *lower* overlap compared to User-initiated Loci (Mdn = 7.00, M = 6.68, SD = 3.34), $t(240) = -2.89$, $p < 0.01$, $r = 0.183$, $\Delta\text{EMM} = -25.8$.

This is qualified by a significant Locus × Structure interaction, F(1, 240) = 5.38, $p = 0.021$, $\eta_p^2 = 0.022$. Post hoc simple-effects analyses revealed a significant difference between LLM- and User-initiated under the Collaborative structure. LLM-Collaborative (Mdn = 6.00, M = 5.99, SD = 2.82) reported significantly *greater* overlap compared to User-Collaborative (Mdn = 6.00, M

= 6.25, SD = 3.40), $t(240) = 2.02$, $p = 0.044$, $r = 0.130$, $\Delta$EMM = 26.1. No other comparisons were significant. See Figure 7-5 for comparison of EMMs.

### 7.4.3 System and Attitudes

This section reports analyses on system-level metrics of time, conversation turns, and request resources, as well as attitudinal metrics towards the intervention through the APOI and SUS. The variables in this section are external to the path model due to either a lack of theoretical relevance or poor fit in initial model tests. For context on the interaction behaviors within the IV structure, a brief breakdown of the proportion of actions taken in each condition, based on their affordances (User Select, User Approve), is provided in Table 7-4.

Table 7-4. Summary of the proportion of **Focus**, **Strategy**, **Utterance**, and **Macro** (average across all steps) actions taken across each condition. For example, within **LLM-Collaborative**, 6% of **Utterances** were selected (i.e., edited or re-suggested), and 94% employed the LLM verbatim suggestion. **User-Solo** and **LLM-Solo** always required the user to Select or Approve respectively.

| Condition Actions | | Focus | Strategy | Utterance | Macro |
|---|---|---|---|---|---|
| **User-Solo** | Select | 1.00 | 1.00 | 1.00 | 1.00 |
| | Approve | 0.00 | 0.00 | 0.00 | 0.00 |
| **User-Collab** | Select | 0.16 | 0.29 | 0.57 | 0.34 |
| | Approve | 0.84 | 0.71 | 0.43 | 0.66 |
| **LLM-Collab** | Select | 0.06 | 0.04 | 0.06 | 0.06 |
| | Approve | 0.94 | 0.96 | 0.94 | 0.94 |
| **LLM-Solo** | Select | 0.00 | 0.00 | 0.00 | 0.00 |
| | Approve | 1.00 | 1.00 | 1.00 | 1.00 |

**Time**. To improve transparency of the descriptives for time-related effects, outliers (n = 18) were removed for time *only*, using a 1.5 interquartile range criterion on the log-transformed variable (results were unchanged). ART ANOVA revealed a significant main effect of Locus on time, F(1, 222) = 16.6, $p < 0.001$, $\eta_p^2 = 0.070$. Regardless of Structure, post hoc analyses revealed that LLM-initiated Loci (Mdn = 37.5, M = 42.3, SD = 17.4) spent *less* time in the intervention than User-initiated Loci (Mdn = 44.3, M = 48.0, SD = 15.6), $t(222) = -4.07$, $p < 0.001$, $r = 0.264$. No other significant effects were found.

**Conversation Turns**. ART ANOVA revealed no significant effects of Locus, Structure, or their interaction on conversation turns ($p > 0.05$).

**Requested Resources**. ART ANOVA revealed a significant main effect of Locus on the quantity of requested resources, F(1, 240) = 6.56, $p = 0.011$, $\eta_p^2 = 0.027$. Regardless of Structure, post hoc analyses revealed that LLM-initiated Loci (Mdn = 0.992, M = 1.00, SD = 1.19) interacted with *greater* quantities of requested resources than User-initiated Loci (Mdn = 0.000, M = 0.607, SD = 0.871), $t(240) = 2.56$, $p = 0.011$, $r = 0.163$. No other significant effects were found.

**APOI**. For skepticism, confidence, and technologization, ART ANOVA revealed no significant effects of Locus, Structure, or their interaction ($p > 0.05$). For anonymity benefits, ART ANOVA revealed a significant Locus × Structure interaction effect F(1, 240) = 5.84, $p = 0.016$, $\eta_p^2 = 0.024$. Post hoc analyses revealed a *greater* (better) report of Anonymity benefits in LLM-Collaborative (Mdn = 4.50, M = 4.28, SD = 1.12) compared to User-Collaborative (Mdn = 3.75, M = 3.70, SD = 1.33), $t(240) = 2.83$, $p = 0.030$, $r = 0.180$ (see Figure 7-5). No other comparisons were significant.

**System Usability**. ART ANOVA revealed a significant Locus × Structure interaction effect on SUS scores F(1, 240) = 5.03, $p = 0.026$, $\eta_p^2 = 0.021$. However, post hoc analyses revealed no significant differences.

## 7.5 Discussion

### 7.5.1 Summary of Key Results

In this study, increasing LLM involvement in the perspective-taking process produced distinct effects across perspective overlap, conversational engagement, and readiness for change.

**$\mathbf{RQ1}_{\mathbf{PerspectiveOverlap}}$**

LLM involvement significantly influenced perspective overlap: LLM-initiated Loci led to *reduced* overlap overall compared to User-initiated Loci, but this main effect is qualified by a significant Locus × Structure interaction. In Collaborative Structures, LLM-initiated Loci led to *greater* overlap compared to User-initiated Loci (LLM-Collaborative > User-Collaborative).

**RQ2$_{\text{ConversationalEngagement}}$**

Overlap positively influenced the rate of change talk, and Collaborative Structure conversations (compared to Solo) also further increased the rate of change talk across conversations. Locus (-) and Locus × Structure (+) also demonstrated significant *indirect* effects on change talk.

**RQ3$_{\text{Readiness}}$**

LLM involvement did not exert direct effects on Δ Readiness; however, the resulting change talk positively influenced Δ Readiness. Structure (+), Locus (-), the Locus × Structure Interaction (+), and overlap (+) also demonstrated significant *indirect* effects on Δ Readiness.

These findings are discussed in the following section alongside broader HCI implications, with additional context provided from system- and intervention-level outcomes.

### 7.5.2 LLMs to Scaffold Perspective-Taking

The findings of this work suggest that LLMs may assist with perspective-taking, but a proper balance between user and LLM involvement is necessary. The identified higher-order interaction effect shows that shifting cognitive tasks from the user to the LLM can improve perspective overlap within collaborative interactions that preserve user autonomy (LLM-Collaborative > User-Collaborative). These findings suggest that when LLM is involved in perspective-taking, allocating demand to LLM can establish – and, in some cases, improve – perspective overlap, as long as user autonomy and oversight are maintained. This implication is particularly relevant given that perspective-taking can be a cognitively demanding process [5, 4]. Perspective-taking depends upon both information about the perspective and motivation to adopt it [32, 147, 187]. Research has shown that perspective-takers rely on a mix of inferential and information-cultivating strategies that can be affected by factors such as cognitive load, unfamiliarity, or lack of energy [146, 147]. Accordingly, this study aimed to facilitate cognitive mechanisms of perspective-taking by employing LLMs to scaffold the acquisition of inferences and information about the counselor's perspective [380]. This approach is underscored by prior work that repeatedly demonstrates AI's capability to reduce extraneous cognitive load [151, 114], scaffold engagement with unfamiliar or complex material [252, 111], and preserve meaningful

user involvement while supporting higher-order, metacognitive processes [41, 52]. While factors like cognitive load are unmeasured in Study 3, the findings of greater time in User-initiated Loci (paired with the notable lack of difference in quantity of conversation turns) would suggest that the temporal demand to craft a counselor-response is reduced by greater LLMs.

Furthermore, these results indicate that greater manual user involvement may not necessarily facilitate greater perspective overlap. Because conditions manifested affordances differently, statistical comparisons of Select and Approve across conditions are not readily interpretable; however, descriptive interaction metrics indicate that LLM-Collaborative users largely defaulted to the LLM's suggestions, manually selecting or editing an action in only 6% of interactions (compared to 34% in User-Collaborative; Table 7-4). Seemingly, the ability to oversee, hear the reasoning behind, and optionally modify the response (LLM-Collaborative) was more effective in producing Overlap than affording greater User-initiative and procedural, metacognitive coaching (User-Collaborative).

### 7.5.3 LLM-Scaffolded Perspective-Taking on Conversational Engagement

This study demonstrates that perspective overlap had a direct, positive effect on the rate of change talk, indicating that greater alignment with the taken counselor perspective corresponded with more frequent production of change-aligned self-responses. Consistent with this pattern, both Locus and the Locus × Structure Interaction influenced change talk indirectly through their effects on overlap (Table 7-3), suggesting that overlap functions as a key mechanism through which LLMs shape conversational engagement. These findings align with extensive literature demonstrating that perspective-taking can alter behavioral engagement [301, 400, 4, 212, 224]. They also reinforce Study 2's findings and related work that perspective-taking can influence conversational engagement, specifically [417, 50].

Beyond the effects of overlap, two notable effects of Structure and Locus are discussed. First, the direct effect of a Collaborative Structure on change talk suggests that collaborative response construction supports change-aligned engagement beyond the effects of perspective-taking alone. This interpretation is consistent with motivational interviewing principles, which emphasize that

change is most likely when clients actively participate in shaping the direction and content of the conversation, rather than passively receiving guidance [279, 183, 333]. This also adds validation to the study's design, as prior meta-analyses demonstrate that proper use of motivational interviewing strategies, including OARS, is associated with increased change talk during counseling interactions [259]. Similar phenomena have also been evidenced in human-AI collaborations, demonstrating that mixed interaction structures and HITL frameworks can promote engagement and behavior change in health contexts [319, 214, 383].

Second, collapsed over Structure, LLM-initiated Loci generally reduced overlap and exerted a negative indirect effect on change talk. However, reduced change talk does not necessarily imply a uniform reduction in conversational engagement. A significant main effect on the quantity of requested resources demonstrates that users in LLM-initiated Loci, compared to User-initiated Loci, requested more information, a form of engagement not reflected in conversation turns or change talk. This may be a related effect of perspective-taking, as perspective-taking has been shown to improve engagement with educational resources [301]. Another possibility is that it is instead a byproduct of the aforementioned reduced time spent within LLM-initiated Loci. In the latter case, it is plausible that LLM involvement may have simply reduced temporal/cognitive demands in crafting counselor-responses and enabled greater capacity to explore supplemental resources [151]. Regardless, these findings highlight a potential tradeoff between different levels of User and LLM initiative. Conversational designs that involve users more directly in the interaction process can promote perspective-aligned engagement (User-initiated Loci), but they may also impose greater demands that *limit* engagement with other aspects of the intervention. In light of the prior discussion on overlap, this tradeoff motivates further consideration of designs that balance LLM scaffolding with user involvement to support multiple dimensions of conversational engagement while moderating the demand on users.

### 7.5.4 Readiness and Attitudes

In this study, change talk emerged as the sole direct influence on improvements in $\Delta$ Readiness. An effect of greater perceived anonymity benefits emerged within the

Collaborative-Structure for LLM-initiated Loci compared to User-initiated Loci (LLM-Collaborative > User-Collaborative). These results extend and clarify findings from Study 2 and prior work in three novel ways.

First, the present study provides further evidence that perspective-taking (overlap) is not a standalone effect on conversational engagement (change talk), and instead, perspective-taking can amplify direct effects on readiness for change. In Study 2, readiness improved across conditions, but the relationship between perspective-taking, conversational engagement, and overlap was not captured. The present study demonstrates that perspective-taking interventions can foster conversational engagements that meaningfully affect changes in readiness for mental health. This also builds upon similar literature indicating that perspective-taking can improve readiness to take pro-social action [380], in the forms of donating money [84], community outreach [34], and volunteering [260]. Worth noting is that no direct effect of the IVs or overlap on readiness was observed; instead, they were significantly mediated through change talk. The absence of direct effects is plausibly attributed to the study's design, in which all participants – regardless of condition – engaged in a motivational interview aimed at supporting reflection and identifying next steps for change. Still, the observed effect of change talk on $\Delta$ Readiness was highly expected, given its grounding in motivational interviewing theory [331] as well as prior empirical work validating the relationship of change talk on behavior adoption or cessation [257, 258]. The observed indirect effects are consistent with the understanding that readiness is supported by change-focused discussion and planning, rather than manipulations in interaction structure.

Second, the present study also found a significantly better perception of anonymity benefits Perceived anonymity benefits reflect the extent to which users feel the intervention was more discrete and less stigmatizing than traditional in-person services [363, 342]. Known effects of perspective-taking on egocentrism, paired with the findings on overlap, may help explain the observed effect. A prominent effect of perspective-taking is its ability to reduce egocentric bias and self-consciousness [178, 366, 340, 122]. Research on computer-mediated communication and perspective-taking has also demonstrated that greater dimensions of self-consciousness are

associated with reduced feelings of anonymity [266, 136, 352, 203]. The comparatively lower overlap observed in User-Collaborative participants may suggest a greater sense of self and a lower sense of anonymity. Despite evidence that computers and VHs are generally perceived as more anonymous than humans [397, 249], the present findings suggest that anonymity perceptions may still be sensitive to how user involvement is structured within the interaction.

Finally, the present findings also build upon an identified limitation of Study 2. The Study 2 intervention employed perspective-taking exclusively without grounding in the user's self-perspective. This manipulation led to greater skepticism, or worsened perceptions that the intervention was personally relevant and effective (see Section 6.4.4). The present study was designed to use a motivational interview in which users alternated between their self-perspective (self-response) and the taken perspective (counselor-response). The present study suggests that perspective-taking may be integratable alongside standard self-perspective engagements to promote self-other overlap. Prior research has demonstrated that perspective-taking is dynamically switchable, especially when switching back to the self-perspective [54, 53]. This has also been explored in HCI and VR contexts, where providing alternating self- and other-perspectives can improve self-attitudes [184, 124]. Therefore, this work demonstrates the potential of structuring VH conversations around alternating self- and other-perspectives to promote the positive effects of perspective-taking while grounding the user.

### 7.5.5 Broader HCI Implications

A primary implication of this work is the ability to use LLMs within HITL frameworks to support perspective-taking in VH conversations. In HCI literature, perspective-taking has often been supported by providing avatars that pair imaginative perspective-taking with visual stimuli [5, 295, 182]. The Pre-Study suggested that including an avatar within VH conversations may have a modest effect on users' perspective change. In contrast, the present study demonstrates that LLMs reduce aspects of perspective-taking demand by adopting an initiating, collaborative role, while still achieving overlap that promotes conversational engagement. This study also provides evidence of the *cost* of perspective-taking. User-initiated Loci within the HITL framework were

generally associated with greater overlap, but came at a cost of greater temporal demand, diminished engagement with system resources, and potential effects on user attitudes. These tradeoffs raise questions about whether the incremental gains in Overlap warrant the additional user burden. Therefore, LLMs within collaborative settings may represent a practical approach to promote perspective-taking in VH conversations without overencumbering users.

Beyond perspective-taking, this research demonstrates that HITL frameworks may be an effective tool for promoting engagement and well-being in VH conversations. When controlling for overlap, human-AI collaboration (LLM-Collaborative) had a positive direct effect on change talk rates and a positive indirect effect on $\Delta$ Readiness. Incorporating structured collaboration between conversational AI systems and users may therefore help orient users towards change. This is supported by prior HCI literature demonstrating that AI-user collaboration can help drive health outcomes [319, 214, 383]. In conversational AI, prior work has integrated HITL frameworks to promote conversational empathy, improve response precision, and personalize educational content [374, 227, 196, 356, 369]. User-state modeling via the classifier also provides transparency in the system's interpretation of users' motivational language explicit within the interaction. HITL approaches can increase accountability and transparency in decision-making, thereby supporting trust and more effective human-AI collaboration [227]. HITL frameworks may be particularly valuable in conversational contexts that depend on partnership between users and counselors, such as motivational interviewing [369]. By distributing initiative between the user and LLM, HITL structures may be able to preserve the collaborative spirit of these conversations while shaping dialogue toward change.

### 7.6 Dissertation Implications

Study 3 directly addresses all three overarching **RQs** in this dissertation. First, Study 3 addresses **RQ1: PerspectiveEngagement** by illustrating how perspective-taking directly influences the rate of change talk. This finding demonstrates that perspective-taking can be selectively positioned to influence specific forms of conversational engagement. Study 3 also demonstrates that the effects of perspective-taking on engagement are not uniformly positive.

Study 3 indicates that the demands of perspective-taking may impact capacity for other interface features, potentially limiting engagement with peripheral supports. Investigation of **RQ2: PerspectiveMentalHealth** is furthered through the findings on readiness. Notably, the change talk produced in this perspective-taking intervention exerted direct positive effects on readiness, and perspective overlap exerted similar indirect effects through change talk. The findings continue to suggest that perspective-taking interventions can be beneficial to self-mental health outcomes.

Finally, this work contributes to **RQ3: PerspectiveAI**, demonstrating opportunities for AI to support the perspective-taking process. This study examined four levels of LLM involvement based on HITL frameworks, ranging from no LLM involvement to collaborative involvement to fully LLM-driven. The findings suggest that LLMs may not necessarily increase overlap, but it seems to improve the process of perspective-taking while establishing overlap. The User-Solo condition functioned as a control without LLM involvement, inherently requiring the greatest user involvement and offering only static (non-LLM) coaching and guidance. As a result, LLM involvement did not produce significantly different levels of overlap compared to fully user-led perspective-taking (User-Solo). However, LLM involvement was seemingly most effective when it was primarily LLM-led and collaborative. Interestingly, users in this group primarily defaulted to the LLM's suggestions rather than asserting their own modifications. This suggests that greater user involvement alone may not be necessary to achieve higher overlap; instead, providing users with transparency and oversight of the interaction may be sufficient to foster a sense of overlap. Given that perspective-taking may not require continuous user demand, an important direction for future work is determining how much of it must be sustained throughout an intervention. It remains unclear whether perspective shifts must occur uniformly across the interaction or whether strategically timed shifts could elicit similar effects. Continuously requiring users to alternate perspectives may impose unnecessary cognitive burden [54, 53]; instead, early orientation to the counselor's perspective may be sufficient to shape the trajectory of the conversation.

## 7.7 Limitations

Three notable limitations are described to contextualize the findings of this work. First, this study investigated the effects of scaffolding within perspective-taking interventions. While User-Solo served as a control for LLM involvement, no formal control for perspective-taking was provided overall. Accordingly, LLM manipulations are primarily evaluated based on notable differences to User-Solo, alongside pairwise comparisons. Second, the User-Solo condition reflects the absence of *LLM*-scaffolded perspective-taking, but not the absence of scaffolded perspective-taking altogether. Participants in this condition still received static coaching on how to select a Focus, Strategy, and Utterance, grounded in motivational interviewing principles and accompanied by static examples. These supports were provided because this dissertation recruited participants from computer science populations, and it was accordingly anticipated that these participants might struggle to adopt the counselor's perspective in a motivational interview. Although numerous measures were taken to ensure participants could complete the intervention (e.g., interactive tutorials, decomposition of response construction into three streamlined steps), the goal was not to evaluate adherence to motivational interviewing techniques. Rather, the intervention aimed to orient participants toward a counselor perspective, allowing flexibility in response formulation and prioritizing perspective enactment over technical correctness. Finally, the results should be interpreted in light of the proportion of variance $R^2$ explained by the path model. The $R^2$ values were modest for overlap, change talk, and readiness, indicating that the specified predictors account for only a limited share of variability in these outcomes and that substantial variance remains unexplained. This suggests that additional individual, contextual, or interaction factors likely contribute to conversational coordination and readiness change beyond the mechanisms tested here. For instance, pre-intervention readiness accounts for a considerable portion of post-intervention variance, but the present model centers on $\Delta$ Readiness to isolate change processes and preserve theoretical clarity and parsimony. Taken together, the findings indicate that the hypothesized pathway from experimental condition to readiness, via perspective

overlap and change talk, is detectable in the data but represents only one component of a broader and multifactorial process of motivational change.