Dual Controller Architecture to Enable Adherence of LLM-based Counselors to Session- and Discourse-level Aspects of Motivational Interviewing Theory

Yermakhan Magzym[1] **April Idalski Carcone**[3] **Elizabeth Towner**[3]
Christopher Mann[4] **M. Safwan Badr**[4] **Alexander Kotov**[1,2]
[1]Department of Computer Science [2]Institute for AI and Data Science
[3]Department of Family Medicine and Public Health Sciences
[4]Department of Internal Medicine
Wayne State University, USA
{yermakhan, kotov}@wayne.edu
{acarcone, ekuhl, cmann, sbadr}@med.wayne.edu

Abstract

Motivational Interviewing (MI) theory posits that effective health behavior counseling requires adherence to a session structure that sequentially progresses through four fundamental processes of MI: engaging, focusing, evoking, and planning. While previous research has demonstrated remarkable capabilities of Large Language Model (LLM)-based conversational agents for MI counseling at mimicking individual utterances of human counselors, methods to ensure adherence of such agents to session- and discourse-level aspects of MI theory have yet to be explored. To address this limitation, we propose the Dual Controller Architecture (DCA), a neuro-symbolic approach to modeling counseling session discourse and structure for LLM-based MI agents with local and global controllers. The global controller ensures timely and appropriate transitions of MI sessions between the stages focused on the four processes. The local controller leverages expert-curated, process-specific distributions of MI counselor communication behaviors to align the agent's responses with the objectives of each process and ensures the agent's utilization of effective counselor communication patterns. Standard approaches to domain adaptation of LLMs, such as prompting, fine-tuning, and RAG, as well as the DCA have been implemented as the five versions of Neural Agent for Obesity MI (NAOMI) and evaluated in a user study involving 49 participants with clinically documented obesity. The study revealed that the DCA demonstrates comparable or superior performance, with particularly strong gains in participants' perceived readiness, importance, and confidence for weight loss.

1 Introduction

Motivational interviewing (MI) (Miller and Rollnick, 2023) is a goal-oriented and client-centered counseling style aimed at building up intrinsic motivation to change through supportive, non-confrontational conversation. Unlike unstructured open-ended social dialogue and structured interactions aimed at completing a certain task, MI theory (Miller and Rollnick, 2023) describes effective counseling sessions as progressing through four fundamental processes of MI: engaging, focusing, evoking, and planning. To form a strong and lasting foundation for sustained efforts towards behavior change, MI counselors guide their clients through these processes, which have distinct focus and goals, using process-specific combinations of communication strategies. MI sessions typically begin with *engaging*, where the counselors aim to develop rapport and a mutually respectful and trusting relationship with the client through empathy and active listening, while helping the client articulate their underlying motivation for behavior change and emphasizing their autonomy. The sessions then proceed to *focusing*, where the counselor works with the client to identify a specific and achievable behavior change goal through open-ended questions and reflections. In *evoking*, the counselor guides the client toward clarifying their own rationale and importance of the behavior change goal identified during focusing with open- and closed-ended questions and reflections. The session concludes with *planning*, where the counselor helps the client develop a plan with clear and actionable steps toward achieving the behavior change goal, with the counselor using a combination of reflections, questions and affirmations to build confidence in adhering to the developed plan.

Recent studies of LLM-based conversational agents for MI counseling, such as MIBot v6 (Mahmood et al., 2025), GPTCoach (Jörke et al., 2025), and Dr. Anderson (Steenstra et al., 2024), demon-

strated the feasibility of fully unscripted automated MI counseling based on zero- or few-shot prompting of pre-trained, closed-source generative LLMs, such as GPT-4. These automated counselors generate their responses based only on conversational context and do not account for session- and discourse-level aspects of MI theory, yielding human counselor-like and MI-adherent responses, but having no control over the structure and progression of an MI session.

To address this limitation, we propose the Dual-Controller Architecture (DCA), a neuro-symbolic approach to modeling session discourse and structure for LLM-based MI counseling agents. DCA provides "scaffolding" for the LLM backbone of these agents, ensuring that they adhere to the aspects of MI theory beyond individual utterances: proper counseling session structure, correct focus of each session stage and utilization of evidence-based discourse-level communication strategies. Specifically, DCA consists of a *global controller*, which ensures that MI agents structure sessions into stages that engage the four MI processes, and a *local controller*, which ensures that MI agents adhere to the focus and objectives of each stage, modeled as stage-specific distributions of counselor communication behaviors, and the utilization of empirically effective multi-turn counselor communication strategies, such as giving advice or information only with the client's permission.

This work contributes to the research on LLM-based conversational agents for MI-based health behavior counseling in the following ways:

- **Modeling the proper structure of MI sessions**: DCA is the first approach to ensure that LLM-based MI agents structure counseling sessions in accordance with MI theory, while preserving the conversational freedom of human-led sessions;
- **Multi-aspect conditioning of agent's responses**: DCA's local controller utilizes a ranking function to first select the next counselor communication behavior code based on MI expert-curated stage-specific distributions and patterns of communication behaviors, as well as a learned whole-session prior, and then steers the backbone LLM toward producing a response grounded not only in MI session context, but also in the session- and discourse-level aspects of MI theory;
- **Rigorous evaluation protocol**: DCA and traditional approaches to domain adaptation of LLMs were evaluated in a user study involving participants referred by primary care clinicians, using validated instruments and human coders trained in assessment of MI fidelity.

2 Related Work

With the emergence of LLMs, the attention of researchers investigating their applications to MI initially focused on methods for specific MI tasks, and later shifted to using them in conversational agents for MI.

2.1 LLM-based Methods for MI-Related Tasks

Research on LLM-based methods for MI tasks has focused on three main directions: the targeted generation of specific MI counselor communication behaviors (primarily reflections and empathy), assessing and improving the quality of counselor responses, and MI session analysis. Specifically, Qian et al. (2023) proposed a method for empathetic response generation based on prompting an LLM that utilized RAG from a corpus of examples and a knowledge base. Min et al. (2024) proposed a multi-armed bandit method to combine the functions measuring reflection score, fluency, and coherence into a unified reward for training an LLM to generate reflections, one of the key MI counselor communication behaviors. Human MI counselor response assessment methods include max-margin ranking for reflection scoring (Min et al., 2022), while the template-based rewriting methods have been proposed to improve the quality of MI reflections (Min et al., 2023) and advice (Welivita and Pu, 2023). A reinforcement learning method (Sharma et al., 2021) has been proposed for rewriting counselor utterances to increase the level of empathy.

Methods for MI session analysis include prompting an LLM to recommend the next counselor behavior code based on the session context (Sun et al., 2025) and a multi-LLM framework for inferring MI strategy rules in natural language from session transcripts (Xie et al., 2024). In contrast to these approaches, our work focuses on leveraging expert-curated distributions of counselor communication behaviors and communication patterns to steer LLM-based counselors towards proper MI session structure and utilization of effective discourse-level communication strategies.

2.2 LLM-based Conversational Agents for Motivational Interviewing

Prior research on LLM-based conversational agents for MI concentrated on two main directions: agents

that simulate clients for training human counselors (Kim et al., 2025; Yang et al., 2025b; Yosef et al., 2024) and agents that simulate MI counselors.

Early conversational agents simulating MI counselors relied on rules and scripted interview scenarios, using intent classifiers and templates to generate specific counselor behaviors such as reflections or affirmations (Park et al., 2019; He et al., 2022; Hua et al., 2025). Later, hybrid systems combined scripts with generative models to improve naturalness of individual responses. For example, *MIBot* v5 (Brown et al., 2023) paired an interview script with an LLM to generate reflections. However, the dependence on hard-coded scripts fundamentally limits conversational flexibility of these agents and is in sharp contrast to the organic flow of MI sessions led by human counselors.

Building on the success of LLMs, recent work has demonstrated the feasibility of using them in fully automated MI counselors. *MIBot* v6 (Mahmood et al., 2025) is a recently proposed automated MI counselor for smoking cessation based on prompting GPT-4o (OpenAI, 2024). Although the evaluation of this agent demonstrated MI adherence of 98% of its utterances, the authors acknowledge critical limitations in their architecture. Specifically, they point out that their adherence metrics "overlook the temporal dynamics of interaction", effectively treating the session as a set of isolated utterances rather than a structured progression. *MIBot* v6 has no session flow control mechanisms other than "reactive oversight", where an external classifier flags deviation from the topic of smoking, after which the session is terminated and removed from analysis. Its evaluation focused on changes in motivation, empathy, and MI fidelity, with fidelity coding performed by an LLM. *GPTCoach* (Jörke et al., 2025), an LLM-based conversational agent to facilitate onboarding to a health coaching program, utilizes prompt chaining to condition counselor responses on data from wearable devices. Although this agent's evaluation based on Motivational Interviewing Treatment Integrity (MITI; Moyers et al., 2016) coding scheme demonstrated high MI fidelity of its individual responses in a controlled setting, it relies on prompt-based responses to track completion of onboarding steps rather than managing a counseling session. Although evaluation of Dr. Anderson (Steenstra et al., 2024), an automated MI counselor for alcohol use based on prompting GPT-4, reveals high levels of empathy of the agent's individual responses, the authors pointed out its architectural limitations, noting that it struggled to "tailor responses to a client's specific stage of change" and failed to "guide conversations toward actionable recovery plans". These limitations align with broader findings in the field: Meyer and Elsweiler (2024) and Gabriel et al. (2024) report that unguided GPT-4 responses often exhibit high variability and uneven MI fidelity.

Taking a client-centric approach, Yang et al. (2025a) proposed CAMI, a general-purpose MI agent. CAMI prompts an LLM to infer the client's mental state and utilizes a 3-level tree to explore topics during a counseling session. However, CAMI conditions the counselor's response on the inferred client mental state, session context, current topic, and behavior code generated based on session context by an LLM, and therefore does not attempt to shape the session structure or ensure utilization of MI best practices at the discourse level. The authors admit that their architecture struggles with "responses being too lengthy and using too many questions", which is indicative of an agent that can maintain an engaging discussion but lacks the mechanisms to structure it toward lasting behavior change. Furthermore, CAMI was evaluated via simulation (automated counselor vs. simulated client) with MITI fidelity coding performed by an LLM, leaving its efficacy with real clients and expert assessment of fidelity unclear.

Other work pointed out further limitations of existing LLM-based counselors. Gabriel et al. (2024) found that GPT-4 responses demonstrated inconsistent MI fidelity and empathy, especially for minority clients, while Karve et al. (2025) noted that although LLM-based MI agents are increasingly feasible, few have been evaluated for long-term behavioral outcomes or adherence to MI theory.

3 Dual-Controller Architecture

The proposed DCA enables LLM-based MI counselors to structure sessions around the four processes of MI and generate responses that align not only with the dialog context, but also with the focus and objectives of each process and empirically effective discourse-level communication patterns. DCA consists of the global and local controllers.

3.1 Global Controller

The global controller illustrated in Figure 1 models an MI session as a sequential progression through the stages corresponding to the four processes of MI: *engaging*, *focusing*, *evoking*, and *planning*.

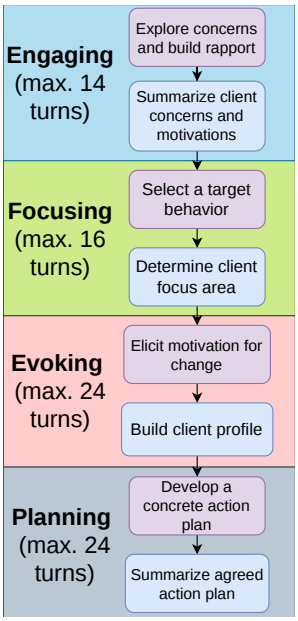

Figure 1: Structure of a weight loss MI session imposed by the DCA's global controller.

Decisions to stay in the current session stage or transition to the stage corresponding to the next MI process are made by the backbone LLM using the classification prompts in Appendix A.2.2. The global controller also transitions to the next stage when a process-specific maximum number of turns is reached. If at this point the agent is in the middle of a pattern (e.g., explicitly requesting client's permission before providing information) or a client poses a question in the immediately preceding turn, the transition is deferred until the pattern is fully completed or the question is answered. Prior to transitioning to the stage corresponding to the next process, a summary of the current stage is generated using the prompt in Appendix A.2.1.

Auxiliary components of the global controller, which use the backbone LLM along with the prompts in Appendix A.2.4, such as the classifier to determine the client's focus area for behavior change during the focusing stage, client profile construction, and creation of summaries at the end of each stage, support agent response generation, but do not affect the session structure.

3.2 Local Controller

At each turn t of the current MI counseling session stage s_t, the local controller first ranks a set of stage-specific candidate counselor codes $\mathcal{C}^{(s_t)}$ (Appendix A.3.1) using the following ranking function based on H_t and $H_t^{(s_t)}$, the code sequences up to turn t in the entire session and the current stage, respectively, and selects the top-ranked code c_t^*:

$$c_t^* = \arg\max_{c_t \in \mathcal{C}^{(s_t)}} \Big[w_{\text{pat}} f_{\text{pat}} \left(c_t, H_t^{(s_t)}, s_t \right) + w_{\text{dist}} f_{\text{dist}} \left(c_t, H_t^{(s_t)}, s_t \right) + w_{\text{gru}} P_\theta(c_t \mid H_t) \Big]$$

where f_{pat} is a binary function rewarding the counselor codes that are continuations of expert-curated evidence-based counselor communication patterns (defined in Table 10 of Appendix A.3.4).

To ensure that MI agent's utterances in a particular counseling session stage collectively adhere to the objectives of the corresponding MI process, f_{dist} rewards the next counselor code that is moving an empirical distribution of counselor communication behaviors for the current stage closer to an expert-curated target distribution for the corresponding process (Table 9 of Appendix A):

$$f_{\text{dist}} \left(c_t, H_t^{(s_t)}, s_t \right) = D_{\text{KL}} \left(\hat{\pi} \left(H_t^{(s_t)} \right) \parallel \pi^{(s_t)} \right) - D_{\text{KL}} \left(\hat{\pi} \left(H_t^{(s_t)}, c_t \right) \parallel \pi^{(s_t)} \right)$$

where $\hat{\pi} \left(H_t^{(s_t)} \right)$ and $\hat{\pi} \left(H_t^{(s_t)}, c_t \right)$ are empirical distributions of counselor communication behaviors for s_t before and after selecting the code c_t for the next agent's utterance, and $\pi^{(s_t)}$ is the target distribution for the MI process corresponding to s_t.

$P_\theta(c_t \mid H_t)$ is a lightweight GRU prior (Appendix H.4) trained on a corpus of MI sessions with human counselors manually annotated with counselor codes (Appendix H.1) to capture short-range dependencies between counselor communication behaviors not covered by the expert-curated patterns. In the deployed system, the weights w_{pat}, w_{dist}, w_{gru} were set to $(0.70, 0.25, 0.05)$, reflecting the a priori preference for expert-curated discourse-level patterns and stage-level focus modeled in the form of distributions of counselor communication behaviors over local counselor code dependencies learned from training corpus (Appendix G.3).

Offline replay ablations (Appendix G.2) indicate that, a posteriori, these three terms have almost equal decision-making power: replay reproduces the controller selections for 90.1% of counselor turns, and removing any one term changes the selected code for 29.2%–35.4% of the turns (Appendix G.1, Appendix G.2). Thus, the local controller needs to take into account MI discourse

patterns, stage-specific communication behavior focus, and a learned prior in order to select the optimal next communication behavior for the agent.

The local controller then prompts Llama-3.1-70B backbone LLM QLoRA-fine-tuned on manually annotated MI session transcripts (Appendix H.1) to generate the counselor's response by providing the description of the selected code c_t^* (see Appendix B.6 for an example). Fine-tuning (Appendix H.3) aimed to expose the backbone LLM to MI code tokens and their descriptions, and to strengthen associations between counselor codes and their textual realizations. Agent's response generation is conditioned on the selected code c_t^*, the current session stage, the cumulative summary of earlier stages, and the context of the current stage, including the client's latest utterance.

The local controller also includes an *elicit-provide-elicit* (EPE) safeguard, a core MI strategy for sharing advice or information while preserving client autonomy. If auxiliary classifiers detect an advice- or information-giving turn and the selected code is GINFO or ADV, the agent first asks permission before providing the content; otherwise, it withholds it and continues the dialogue. This mechanism is intended to improve MI adherence by reducing unsolicited advice.

4 Evaluation

The DCA, along with the standard approaches to domain adaptation of LLMs, such as prompt engineering, fine-tuning and two different variations of retrieval-augmented generation (RAG) (Lewis et al., 2020) have been implemented[1] as the five versions of the **N**eural **A**gent for **O**besity **M**otivational **I**nterviewing (NAOMI), a conversational agent with textual chat-based Web interface (Fig. 13 in Appendix), and comprehensively evaluated in a user study, according to its previously published protocol (Kotov et al., 2024).

4.1 Baselines

The choice of baselines was motivated by the work of Kermani et al. (2025), who conducted a systematic evaluation comparing prompt engineering, fine-tuning, and RAG to identify the optimal deployment strategy of LLMs for mental health-related classification tasks (e.g., detecting depression). We chose `Llama-3.1-70B` (Grattafiori et al., 2024) as the backbone LLM since it is open-weight and could be deployed locally to protect privacy of client data.

[1]source code available at https://github.com/teanalab/NAOMI.

4.1.1 NAOMI-PT: Few-shot Prompt Engineering

To facilitate comparison with existing LLM-based MI counselors, this version relied only on a fixed system prompt (Appendix B) and few-shot examples of key MI counselor communication behaviors such as open- and closed-ended questions, reflections and affirmations (Appendix B.1.2) to guide agent response generation. We iteratively refined the prompt and examples following prompt engineering guidelines (Vatsal and Dubey, 2024) to help the LLM better grasp the nuances of MI counseling.

4.1.2 NAOMI-FT: Fine-tuning

We adapted the backbone LLM to the linguistic style of MI counselors through Supervised Fine-Tuning (SFT) (Ouyang et al., 2022) on our curated corpus of real-life weight loss MI session transcripts with human counselors (Appendix H.1). Since full-parameter fine-tuning incurs significant memory overhead, we used 4-bit quantization with Low-Rank Adaptation (QLoRA) (Dettmers et al., 2023). At inference, NAOMI-FT was also prompted, using the same system prompt as NAOMI-PT, but without the few-shot examples (Appendix B.2). Detailed implementation and training procedures are provided in Appendix H.

4.1.3 NAOMI-RAG: Retrieval-Augmented Generation

NAOMI-RAG uses the same fine-tuned backbone LLM as NAOMI-FT, but conditions the agent response generation on examples of human counselor utterances in similar dialog situations retrieved from the corpus of weight loss MI session transcripts. Specifically, we indexed this corpus with FAISS (Douze et al., 2026) using `nomic-embed-text` embeddings (Nussbaum et al., 2024). At each agent turn, the RAG system encodes the most recent 5-turn session window and retrieves the most similar entries from the indexed corpus. It then extracts only the counselor responses from the retrieved entries and inserts them into the prompt as additional examples for response generation (see Appendix H.1 for corpus statistics and Appendix B.3 for the prompts).

4.1.4 NAOMI-RAG+: Generate-then-revise Approach

NAOMI-RAG+ uses the same fine-tuned backbone LLM as in NAOMI-FT in a two-step pipeline, which resembles *example-based editorial reasoning*: a draft response generated by the backbone LLM is revised based on retrieved human counselor examples and the MI session context. Specifically, in

this approach, a draft response is first generated by the backbone LLM based on the session context alone and the NAOMI-FT prompt. Next, using the same approach as in NAOMI-RAG, the RAG system encodes the most recent 5-turn session window, retrieves similar fragments from the corpus of MI session transcripts, and extracts the counselor responses from those fragments. The same fine-tuned backbone LLM is then prompted for the second time (Appendix B.4) to revise the draft agent response conditioned on the retrieved examples of human counselor responses and the current session context. This approach uses retrieval only during the revision step, with the goal of making the final response better match human MI counselor language and communication behavior.

4.2 User Study

The five versions of NAOMI were evaluated in a study with participants referred from a network of primary care clinics following clinician chart review. Eligible participants were English-speaking adults aged 18–65 with overweight or moderate obesity (Body Mass Index of 25.0–39.9) (additional inclusion and exclusion criteria are provided in Appendix C). Participants had one study visit (Appendix C.1) where they completed demographics and pre-interaction acceptability questionnaires (Tables 11, 12 and 13 in Appendix) along with importance, confidence and readiness rulers (Table 14 in Appendix), interacted with one version of NAOMI (baseline or NAOMI-DCA), then completed post-interaction acceptability, usability and empathy questionnaires (Tables 12, 15, 16 and 17 in Appendix) and had a one-week follow-up where they completed motivational rulers. A total of 49 participants were recruited in the study and were compensated with a $75 gift card for completing the study visit and with a $15 gift card for the one-week follow-up. A comprehensive breakdown of participant demographics, including socioeconomic status, is provided in Appendix C.4.

5 Results

We compared the five versions of NAOMI across three dimensions: 1) participants' perceptions of agent's usability, acceptability and empathy measured by the validated survey instruments (Appendix D.1); 2) short-term change in motivation levels measured by weight loss readiness, importance, and confidence rulers administered pre-, post-, and one week after interaction with each version; 3) MI fidelity assessed based on the Motivational Interviewing Treatment Integrity (MITI) (Moyers et al., 2016) codes assigned to the agent's responses by an expert MITI coder. To assess coding reliability, 20% of sessions were independently co-coded by the second expert MITI coder, revealing high levels of coder agreement (intraclass correlation coefficient (Shrout and Fleiss, 1979) of 0.91 and Cohen's κ (McHugh, 2012) of 0.74). Questionnaire responses of four participants were excluded from data analysis, since they had almost no variance across questions, although session transcripts of those participants were used for MI fidelity assessment (see Appendix E for details).

5.1 Usability

Participants reported generally positive overall experience with all versions of NAOMI (more details in Appendix E). Notably, NAOMI-DCA, our proposed architecture, received the highest ratings for *pleasantness of interaction* (4.38/5), *ease of understanding* (4.50), *supporting self-expression* (4.50), *comfort discussing weight loss* (4.38), and *willingness to use NAOMI again* (4.25). Although NAOMI-RAG and NAOMI-RAG+ scored slightly higher on message-level qualities such as *message form* and *message understandability*, NAOMI-DCA received the strongest ratings on user-centered interaction measures, particularly *supporting self-expression* and *comfort discussing weight loss*.

5.2 Acceptability

Post-interaction acceptability survey responses (more details in Appendix E) showed slightly mixed results. NAOMI-DCA received the highest ratings for *AI agents as acceptable way to receive counseling* (4.12), *saving time relative to in-person counseling* (4.62), *fit with lifestyle* (4.38), and *willingness for long-term use* (4.25). However, NAOMI-PT scored slightly higher on *fit with preferred way to receive counseling* (3.67 vs. 3.62).

5.3 Empathy

Perceived empathy scores summarized in Table 1 that are based on participants' responses to the Consultation and Relational Empathy (CARE) measure (Mercer et al., 2004) indicate that all five versions of NAOMI performed very closely, with NAOMI-RAG receiving the highest mean score of 41.34 and NAOMI-DCA placing in the middle with the mean score of 39.89.

Notably, this ranking does not coincide with the MITI empathy global score (based on the MITI primary coder) in Table 4, for which NAOMI-DCA placed the highest. CARE measure captures the

System	Score (Mean ± SD)	% Perfect
NAOMI-PT	39.22 ± 7.56	11.1
NAOMI-FT	40.00 ± 9.81	20.0
NAOMI-RAG	41.34 ± 4.87	0.0
NAOMI-RAG+	38.43 ± 7.37	14.3
NAOMI-DCA	39.89 ± 6.47	0.0

Table 1: CARE scores by version of NAOMI. % Perfect is the percentage of participants who gave the highest rating for each CARE item.

client's subjective perceptions of counselor's empathic skills, whereas the MITI empathy score is intended to capture an expert coder's assessment of all efforts by a counselor to understand the client's perspective and worldview and convey that understanding to the client.

5.4 Changes in Motivation Levels

Durable positive effect on behavior change precursors (importance, confidence, and readiness) is one of the main goals of MI counseling. Figure 2 illustrates the effect of weight loss MI counseling by different versions of NAOMI on short- and long-term changes in participants' self-reported importance, confidence, and readiness to lose weight.

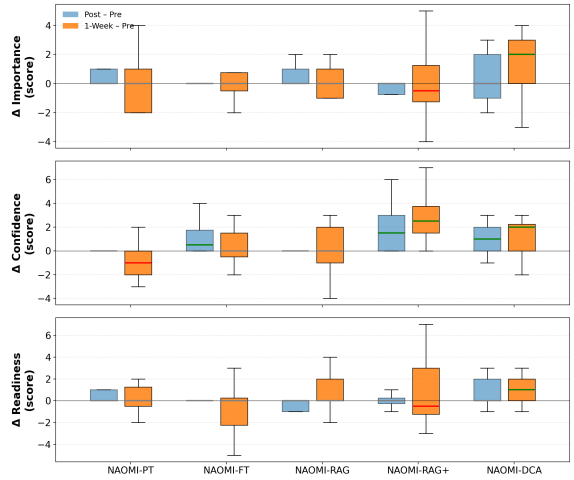

Figure 2: Distribution of changes in behavior change precursors (weight loss importance, confidence and readiness) measured after interaction with different versions of NAOMI and at one-week follow-up relative to pre-interaction baselines. See Table 2 for the p-values of individual comparisons.

Interacting with NAOMI-DCA resulted in consistently positive one-week improvements across all three behavior change precursors (importance Δ of +1.11, confidence Δ of +0.89, and readiness Δ of +0.67). In contrast, the baseline models (PT/FT) yielded only modest or mixed changes (e.g., confidence Δ of -0.88 for PT and readiness Δ of -0.75 for FT). Interacting with NAOMI-RAG+ resulted in the strongest one-week gain in confidence (Δ of +2.22), but smaller gains in other precursors.

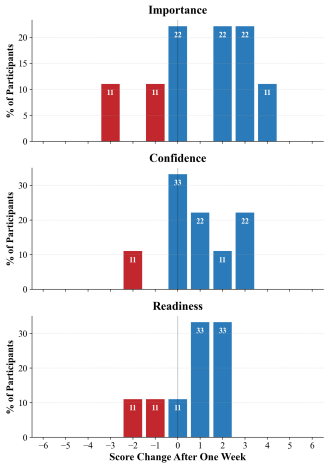

Figure 3: Distribution of one-week changes in behavior change precursors for NAOMI-DCA (see Figure 9 of Appendix E for the same plots corresponding to other versions). Blue bars correspond to positive change or maintenance, red bars to negative change.

Figure 3 illustrates the distributions of one-week changes in the values of behavior change precursors for NAOMI-DCA. Participants largely reported an increase for all three precursors: non-negative change in importance, confidence and readiness was observed for 77.8%, 88.9% and 77.8% of participants, respectively (with gains concentrated at +1 and +2). These observations suggest that the adherence to session- and discourse-level aspects of MI theory enabled by the DCA, particularly ensuring that MI sessions include the *planning* stage, where an agent breaks down weight loss into clear and actionable steps, may be particularly helpful for boosting all three precursors.

	NAOMI-DCA vs.				
	PT	FT	RAG	RAG+	Power
Post − Pre					
Importance	0.06	0.09	0.09	0.07	0.42
Confidence	0.08	0.12	0.11	0.14	0.36
Readiness	0.08	0.09	0.11	0.08	0.40
Follow-up − Pre					
Importance	**0.05**	0.06	0.08	**0.05**	0.47
Confidence	0.07	**0.05**	0.08	**0.05**	0.47
Readiness	0.09	**0.04**	**0.05**	0.08	0.46

Table 2: Uncorrected p-values indicating statistical significance of pairwise comparisons of changes in motivational precursors between NAOMI-DCA and each baseline based on the two-sided Welch t-test of participant-level change values. Entries, in which $p \leq 0.05$, are bolded. Power values are averaged over all comparisons.

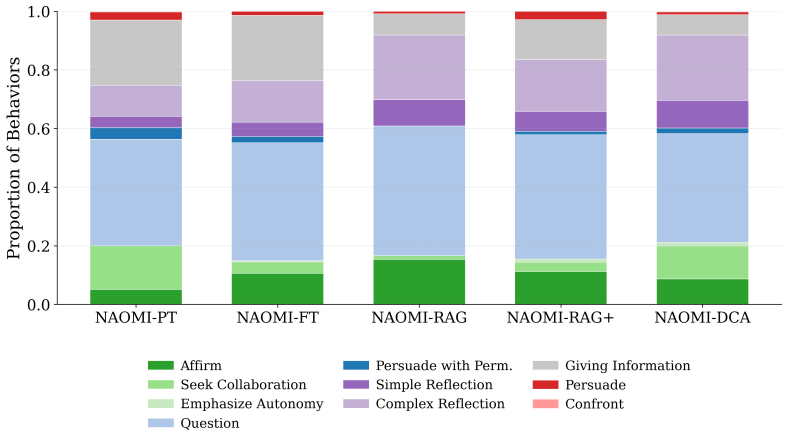

Figure 4: Distribution of MITI codes by the version of NAOMI.

	n	R/Q↑	%CR↑	%MIC↑
Basic	–	≥ 1.0	≥ 0.40	≥ 0.90
Advanced	–	≥ 2.0	≥ 0.50	≥ 1.00
NAOMI-PT	9	0.43	0.68^a	0.89
NAOMI-FT	10	0.50	0.77^a	0.94^b
NAOMI-RAG	10	0.70	0.73^a	0.97^b
NAOMI-RAG+	9	0.59	0.70^a	0.83
NAOMI-DCA	10	0.87	0.65^a	0.95^b

Table 3: MI behavior counts (means across all sessions). ^{a} and ^{b} denote meeting the Advanced and Basic thresholds for MI proficiency, respectively.

	Cultivate↑	Soften↑	Partner↑	Empathy↑
NAOMI Systems				
NAOMI-PT	2.11	3.44	2.11	1.89
NAOMI-FT	2.50	3.30	2.80	2.50
NAOMI-RAG	**3.40**	**3.90**	2.89	2.89
NAOMI-RAG+	2.78	3.33	2.56	2.56
NAOMI-DCA	3.10	3.40	**3.00**	**3.00**

Table 4: MITI global scores for different versions of NAOMI. "Cultivate", "Soften" and "Partner" correspond to Cultivating Change Talk, Softening Sustain Talk, and Partnership, respectively.

Statistical considerations. Following the ORBIT model for developing behavioral interventions (Czajkowski et al., 2015), this study was designed as an early-stage feasibility study, with sample sizes to evaluate each version of NAOMI determined in accordance with the literature on the design of such studies (Bowen et al., 2009). These sample sizes provide sufficient power to detect only large effects. Comparisons of these changes across versions of NAOMI (Figure 2) should be interpreted in accordance with the statistical significance summarized in Table 2. Statistically insignificant comparisons can be interpreted as descriptive trends.

As follows from Table 2, the follow-up − pre-interaction change in at least one precursor in comparisons between DCA and the baselines is statistically significant at 95% confidence level.

5.5 MI Fidelity

Reflection-to-Question ratio (R/Q). A common limitation of LLM-based counselors is question stacking. NAOMI-PT exhibited this clearly with a low R/Q of 0.43. However, NAOMI-DCA substantially improved this ratio to 0.87, outperforming the NAOMI-RAG+ version (0.59) and other baselines. While this still falls short of the basic MI proficiency threshold (1.0), it represents a significant step towards MI expertise, driven by ability of the local controller to cap excessive questioning.

MITI global scores. Table 4 reports the MITI global scores, which consist of expert-rated specific counseling skills, such as Cultivating Change Talk, Softening Sustain Talk, Partnership, and Empathy.

MITI global scores reveal split leadership between NAOMI-RAG and NAOMI-DCA. NAOMI-RAG scored highest on Cultivating Change Talk and Softening Sustain Talk, indicating stronger performance on eliciting client language in favor of change and avoiding reinforcement of sustain talk. However, RAG-based versions showed weaker gains in post-interaction readiness (Section 5.4).

In contrast, NAOMI-DCA achieved the highest Partnership and Empathy scores (3.00 on both), while also maintaining a near-perfect MI consistency (%MIC = 0.95; Table 3). It also remained competitive in terms of Cultivating Change Talk (3.10), despite stricter session- and stage-level constraints. In summary, these results indicate that the DCA demonstrates high levels of MI fidelity while creating strong therapeutic alliance with a client

Version	Mean ± SD	Min	Max
NAOMI-PT	1.35 ± 1.33	0.29	4.90
NAOMI-FT	1.08 ± 1.60	0.14	6.98
NAOMI-RAG	0.98 ± 1.05	0.14	5.89
NAOMI-RAG+	0.72 ± 0.57	0.09	3.29
NAOMI-DCA	**0.43 ± 0.30**	**0.06**	**1.94**

Table 5: KL divergence between session stage-level empirical and target distributions of counselor behaviors, aggregated over all sessions. Lower is better.

through empathy, which we believe translated into effective goal setting and planning and, ultimately, into strong gains in behavior change precursors.

Agent communication behavior profiles. Figure 4 illustrates the distributions of MITI codes assigned to different versions of NAOMI. Rather than exhibiting dominant communication behaviors, NAOMI-DCA employs a balanced combination of questions, reflections, collaboration-seeking, and information provision.

In particular, NAOMI-DCA shows a reduced reliance on *Giving Information* (8.2%) relative to the prompt-based (PT/FT; ≈22%) and the RAG+ baselines (13.7%), while maintaining moderate levels of *Seeking Collaboration* (13.3%) and the highest proportion of Simple (9.35%) and Complex Reflections (22.3%) among all versions of NAOMI. This contrasts with simpler baseline models, which frequently default to information delivery, and with RAG-based approaches, which are dominated by questions and reflections, but less often seek client permission before giving advice or information.

Adherence to MI process goals. To assess adherence to high-level MI process goals rather than the individual turns, we segmented every transcript into the four MI processes and annotated counselor utterances with communication behaviors, applying the same automated procedure to all five versions. For each stage we computed the KL divergence between the empirical distribution of counselor behaviors and the expert-curated target distribution for the corresponding MI process (Table 9 in Appendix). As follows from Table 5, NAOMI-DCA attains the lowest mean divergence and the narrowest spread, with its worst stage-level divergence below the worst session of every baseline.

5.6 Comparison with Other MI Agents

Table 6 shows that NAOMI-DCA achieves the highest %MIC among the open-weight agents and the highest scores on three of four MITI global scores. Its one-week gains in Importance and Readiness (1.11 and 0.67) also exceed those reported for MI-Bot (0.50 and 0.30), which achieves larger confidence gain (1.70 vs. 0.89). These cross-study comparisons are provided for reference only, as there are substantial differences in evaluation protocols. In particular, LLM-simulated clients were used in evaluations of DIIR, CoS, and CAMI, while LLM-based MITI coding was used for CAMI.

	Behavioral			Relational			
Agent	R/Q	%CR	%MIC	Cult.	Soft.	Part.	Emp.
Proprietary LLM (GPT-4 / 4o)							
DIIR	0.42	**82.4**	89.1	2.21	2.64	2.37	3.10
CoS†	0.29	49.1	94.5	2.40	2.71	2.33	3.23
CAMI	0.56	57.3	96.6	**2.62**	**2.86**	**2.58**	**3.37**
Dr. Anderson	**1.86**	63.0	94.0	–	–	–	–
GPTCoach	0.13	50.0	93.3‡	–	–	–	–
MIBot	1.30	–	**98.0**	–	–	–	–
Open-weight LLM (LLaMA 3.1 70B)							
DIIR	0.97	**71.5**	85.4	2.13	2.57	2.28	2.87
CoS†	0.77	58.9	87.1	2.14	2.61	2.13	3.07
CAMI	**1.11**	60.5	90.7	2.38	2.78	2.37	**3.33**
NAOMI-DCA	0.87	65.0	**95.0**	**3.10**	**3.40**	**3.00**	3.00

Table 6: Comparison with published LLM-based MI agents; higher is better for all seven metrics. Values for DIIR, CoS and CAMI are as reported by Yang et al. (2025a); †ablated reimplementation. Values for Dr. Anderson, GPTCoach and MIBot are from Steenstra et al. (2024), Jörke et al. (2025) and Mahmood et al. (2025) respectively; ‡derived from the reported MI-inconsistent response rate. "–" indicates not reported.

6 Conclusion

This paper proposed a dual-controller architecture to ensure that LLM-based counselors move clients through the four processes of MI during counseling sessions, helping clients articulate their underlying motivation for behavior change, collaboratively identifying a behavior change goal, clarifying the client's rationale for the selected goal, and developing action steps toward achieving this goal. The analysis of participant responses in surveys of usability, acceptability and empathy as well as standardized assessment of MI fidelity indicates that the DCA demonstrates comparable or superior results to baseline methods, but greater improvements in participants' perception of weight loss readiness, importance, and confidence. The multi-aspect conditioning of agent responses in the local controller to steer them toward alignment with discourse-level aspects of the dialog is not specific to MI and can be utilized in conversational agents for any domain with limited training data.

7 Limitations

We recognize that the global controller in the proposed DCA architecture is tailored specifically to MI and its process-oriented session structure, and may therefore not be directly applicable to other counseling modalities that do not follow the same progression through session stages. This design choice is justified by the objective of the present work, which is to develop a method to ensure adherence of LLM-based counselors to the session- and discourse-level aspects of MI theory. At the same time, the lack of corpora consisting of transcripts of MI sessions with human counselors that go through all four processes and are annotated with the transition points between them remains an important limitation for future work along this avenue. Availability of such corpora would not only allow learning transitions between the four processes and per-process distributions of counselor communication behaviors directly from data, but also allow the use of offline reinforcement learning methods, such as Implicit Q-Learning (Kostrikov et al., 2022), to train a policy network to generate trajectories of counselor communication behaviors that align with the objectives of each MI process.

The nature of this study also limits its outcomes. The observed changes in importance, confidence, and readiness should be interpreted as descriptive trends rather than as evidence of clinical efficacy. Establishing clinical efficacy of DCA would require a fully powered clinical trial that, in the typical progression of biomedical research, follows feasibility studies like ours. A further consequence of small sample size is that the study cannot attribute the observed gains in precursors of behavior change to any individual component of the proposed architecture. User-facing ablation of DCA components thus remains future work.

The proposed architecture combines several LLM-based components: the backbone LLM generating counselor responses, the classifiers determining stage transitions and client focus, and the summarizers that condense the counselor-client exchanges within each stage. Because the outputs of all these auxiliary components are part of the prompt, an error introduced by any of them can propagate into the agent's responses even when its next communication behavior is chosen correctly. Appendix J.2 documents such a case in our study, where a summary for the planning stage introduced directive phrasing that was carried into the response and identified by MITI coders as persuasion, although no persuasive behavior code had been selected by the local controller. Controlling the response type is therefore not sufficient on its own to ensure MI fidelity. We note, however, that all versions of NAOMI were evaluated end-to-end, with every component operating in tandem during sessions with real participants, so that failures of this kind are reflected in expert MITI coding and in participant survey responses.

Ethical Considerations

We would like to emphasize that the present study should not be interpreted as evidence that LLM-based agents can or should replace human MI counselors. Participants in the user study of different versions of NAOMI were explicitly instructed to continue with their current weight loss treatment and counseling programs (if any) and not view NAOMI as a replacement for licensed human motivational interviewing counselors. Rather, our findings are limited to the evaluation of a particular architecture for structuring LLM-based MI counseling sessions in a controlled research setting. To improve safety, we've integrated Llama Guard (Inan et al., 2023), which monitors both input and output, reducing the risk of unsafe content from being shown to the client. Realizing that such safeguards are inherently imperfect and cannot guarantee that harmful or inappropriate responses will not be generated, all participant interactions with NAOMI were done under the supervision of a study team member. Any real-world therapeutic deployment would require substantially more extensive validation and stronger safety mechanisms.

Acknowledgments

This work was supported by the National Institutes of Health under the award number R21NR020388. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We sincerely thank the anonymous reviewers for their detailed feedback and helpful suggestions.

References

Deborah J. Bowen, Matthew Kreuter, Bonnie Spring, Ludmila Cofta-Woerpel, Laura Linnan, Diane Weiner, Suzanne Bakken, Cecilia Patrick Kaplan, Linda Squiers, Cecilia Fabrizio, and Maria Fernandez. 2009. How we design feasibility studies. *American Journal of Preventive Medicine*, 36(5):452–457.

A Brown, AT Kumar, O Melamed, I Ahmed, YH Wang, A Deza, M Morcos, L Zhu, M Maslej, N Minian,

V Sujaya, J Wolff, O Doggett, M Iantorno, M Ratto, P Selby, and J Rose. 2023. A motivational interviewing chatbot with generative reflections for increasing readiness to quit smoking: Iterative development study. *JMIR Mental Health*, 10:e49132.

April Idalski Carcone, Sylvie Naar-King, Kathryn E. Brogan, Terrance Albrecht, Ellen Barton, Tanina Foster, Tim Martin, and Sharon Marshall. 2013. Provider communication behaviors that predict motivation to change in black adolescents with obesity. *Journal of Developmental and Behavioral Pediatrics*, 34(8):599–608.

Matthew M Carper, R Kathryn McHugh, Heather W Murray, and David H Barlow. 2014. Psychometric analysis of the perceptions of computerized therapy questionnaire-patient version (PCTQ-P). *Administration and Polgicy in Mental Health and Mental Health Services Research*, 41(1):104–113.

Susan M. Czajkowski, Lynda H. Powell, Nancy Adler, Sylvie Naar-King, Kim D. Reynolds, Christine M. Hunter, Barbara Laraia, Deborah H. Olster, Frank M. Perna, Janey C. Peterson, Elissa Epel, Josephine E. Boyington, and Mary E. Charlson. 2015. From ideas to efficacy: The ORBIT model for developing behavioral treatments for chronic diseases. *Health Psychology*, 34(10):971–982.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2026. The Faiss library. *IEEE Transactions on Big Data*, 12(2):346–361.

Saadia Gabriel, Isha Puri, Xuhai Xu, Matteo Malgaroli, and Marzyeh Ghassemi. 2024. Can AI relate: Testing large language model response for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2206–2221, Miami, Florida, USA. Association for Computational Linguistics.

Google DeepMind. 2026. Gemini 3.1 pro model card. Accessed: March 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Linwei He, Erkan Basar, Reinout W. Wiers, Marjolijn L. Antheunis, and Emiel Krahmer. 2022. Can chatbots help to motivate smoking cessation? A study on the effectiveness of motivational interviewing on engagement and therapeutic alliance. *BMC Public Health*, 22(1):726.

Yining Hua, Steve Siddals, Zilin Ma, Isaac Galatzer-Levy, Winna Xia, Christine Hau, Hongbin Na, Matthew Flathers, Jake Linardon, Cyrus Ayubcha, and John Torous. 2025. Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: a systematic review. *World Psychiatry*, 24(3):383–394.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *Preprint*, arXiv:2312.06674.

Matthew Jörke, Shardul Sapkota, Lyndsea Warkenthien, Niklas Vainio, Paul Schmiedmayer, Emma Brunskill, and James A. Landay. 2025. Gptcoach: Towards llm-based physical activity coaching. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA. Association for Computing Machinery.

Zev Karve, Jacob Calpey, Christopher Machado, Michelle Knecht, and Maria Carmenza Mejia. 2025. New doc on the block: Scoping review of ai systems delivering motivational interviewing for health behavior change. *J Med Internet Res*, 27:e78417.

Arshia Kermani, Veronica Perez-Rosas, and Vangelis Metsis. 2025. A systematic evaluation of LLM strategies for mental health text analysis: Fine-tuning vs. prompt engineering vs. RAG. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 172–180, Albuquerque, New Mexico. Association for Computational Linguistics.

Minju Kim, Dongje Yoo, Yeonjun Hwang, Minseok Kang, Namyoung Kim, Minju Gwak, Beong-woo Kwak, Hyungjoo Chae, Harim Kim, Yunjoong Lee, Min Hee Kim, Dayi Jung, Kyong-Mee Chung, and Jinyoung Yeo. 2025. Can you share your story? modeling clients' metacognition and openness for LLM therapist evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25943–25962, Vienna, Austria. Association for Computational Linguistics.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. 2022. Offline reinforcement learning with implicit Q-learning. In *International Conference on Learning Representations*.

Alexander Kotov, April Idalski Carcone, and Elizabeth Towner. 2024. Neural conversational agent for weight loss counseling: protocol for an implementation and feasibility study. *JMIR Research Protocols*, 13(1):e60361.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich

Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, volume 33.

Zafarullah Mahmood, Soliman Ali, Jiading Zhu, Mohamed Abdelwahab, Michelle Yu Collins, Sihan Chen, Yi Cheng Zhao, Jodi Wolff, Osnat C. Melamed, Nadia Minian, Marta Maslej, Carolynne Cooper, Matt Ratto, Peter Selby, and Jonathan Rose. 2025. A fully generative motivational interviewing counsellor chatbot for moving smokers towards the decision to quit. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25008–25043, Vienna, Austria. Association for Computational Linguistics.

Tim Martin, Theresa B. Moyers, Jon M. Houck, Paulette J. Christopher, and William R. Miller. 2005. *Motivational Interviewing Sequential Code for Observing Process Exchanges (MI-SCOPE) Coder's Manual*. Center on Alcoholism, Substance Abuse, and Addictions, University of New Mexico, Albuquerque, NM.

Mary L. McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3):276–282.

Stewart W Mercer, Margaret Maxwell, David Heaney, and Graham CM Watt. 2004. The consultation and relational empathy (care) measure: development and preliminary validation and reliability of an empathy-based consultation process measure. *Family Practice*, 21(6):699–705.

Selina Meyer and David Elsweiler. 2024. "You tell me": A dataset of GPT-4-based behaviour change support conversations. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*, page 411–416, New York, NY, USA. Association for Computing Machinery.

William R. Miller, Theresa B. Moyers, Denise Ernst, and Paul Amrhein. 2003. *Manual for the Motivational Interviewing Skill Code (MISC), Version 2.0*. Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico, Albuquerque, NM.

William R. Miller and Stephen Rollnick. 2023. *Motivational Interviewing: Helping People Change and Grow*, 4th edition. Guilford Press.

Do June Min, Veronica Perez-Rosas, Ken Resnicow, and Rada Mihalcea. 2023. VERVE: Template-based ReflectiVE rewriting for MotiVational IntErviewing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10289–10302, Singapore. Association for Computational Linguistics.

Do June Min, Veronica Perez-Rosas, Ken Resnicow, and Rada Mihalcea. 2024. Dynamic reward adjustment in multi-reward reinforcement learning for counselor reflection generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5437–5449, Torino, Italia. ELRA and ICCL.

Do June Min, Verónica Pérez-Rosas, Kenneth Resnicow, and Rada Mihalcea. 2022. PAIR: Prompt-Aware margIn Ranking for counselor reflection scoring in motivational interviewing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 148–158, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Theresa B. Moyers, Lauren N. Rowell, Jennifer K. Manuel, Denise Ernst, and Jon M. Houck. 2016. The motivational interviewing treatment integrity code (MITI 4): Rationale, preliminary reliability and validity. *Journal of Substance Abuse Treatment*, 65:36–42.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. *Preprint*, arXiv:2402.01613.

OpenAI. 2024. GPT-4o system card.

OpenAI. 2026. GPT-5.4 Thinking system card. Accessed: March 2026.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

S Park, J Choi, S Lee, C Oh, C Kim, S La, J Lee, and B Suh. 2019. Designing a chatbot for a brief motivational interview on stress management: Qualitative case study. *Journal of Medical Internet Research*, 21(4):e12231.

Bambang Parmanto, Al N Lewis Jr, Kelley M Graham, and Marnie H Bertolet. 2016. Development of the telehealth usability questionnaire (TUQ). *International Journal of Telerehabilitation*, 8(1):3–10.

Yushan Qian, Weinan Zhang, and Ting Liu. 2023. Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6516–6528, Singapore. Association for Computational Linguistics.

Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference 2021*, pages 194–205. Association for Computing Machinery.

Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428.

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 24th ACM International Conference on Intelligent Virtual Agents*, New York, NY, USA. Association for Computing Machinery.

Xin Sun, Xiao Tang, Abdallah El Ali, Zhuying Li, Pengjie Ren, Jan de Wit, Jiahuan Pei, and Jos A.Bosch. 2025. Rethinking the alignment of psychotherapy dialogue generation with motivational interviewing strategies. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1983–2002. Association for Computational Linguistics.

Shubham Vatsal and Harsh Dubey. 2024. A survey of prompt engineering methods in large language models for different NLP tasks. *Preprint*, arXiv:2407.12994.

Anuradha Welivita and Pearl Pu. 2023. Boosting distress support dialogue responses with motivational interviewing strategy. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5411–5432, Toronto, Canada. Association for Computational Linguistics.

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.

Yizhe Yang, Palakorn Achananuparp, Heyan Huang, Jing Jiang, Phey Ling Kit, Nicholas Gabriel Lim, Cameron Tan Shi Ern, and Ee-Peng Lim. 2025a. CAMI: A counselor agent supporting motivational interviewing through state inference and topic exploration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21037–21081, Vienna, Austria. Association for Computational Linguistics.

Yizhe Yang, Palakorn Achananuparp, Heyan Huang, Jing Jiang, Nicholas Gabriel Lim, Cameron Tan Shi Ern, Phey Ling Kit, Jenny Giam Xiuhui, John Pinto, and Ee-Peng Lim. 2025b. Consistent client simulation for motivational interviewing-based counseling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20959–20998, Vienna, Austria. Association for Computational Linguistics.

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.

A Details of the Dual Controller Architecture

A.1 Dynamic Prompt

At each turn, DCA constructs the counselor response generation prompt dynamically from the current session state rather than relying on a fixed template. The prompt includes: (i) a fixed counselor persona header, (ii) the policy for the current MI session stage, (iii) session memory, including a cumulative summary of key information from earlier stages and client facts, and (iv) the recent dialogue context for the current stage. The final instruction then appends the client's latest utterance, the selected counselor code, and the definition of that code. Importantly, the implementation provides only the definition of the *selected* code, rather than definitions for all codes. Full prompt templates are provided in Appendix B.5, and an example of counselor code definition is shown in Appendix B.6.

This design makes generation both stage-aware and discourse-aware: stage policies constrain the broader counseling objective, while the selected code definition sharpens the intended communicative function of the next response. We further evaluate this design choice in the dynamic-prompt ablation in Appendix G.4, which compares prompts with no code definitions, all stage-valid code definitions, and only the selected code definition. As shown in Table 19 and Figure 11, including only the selected code definition yields the strongest realization of the intended counselor behavior.

A.2 Details of Global Controller

A.2.1 Stage Summarization Prompt

At the conclusion of each session stage, a separate summarization prompt is triggered to generate a summary sentence to concisely capture the motivational language expressed by the client.

Stage Summary Prompt

You are a summarizer for a single stage of a Motivational Interviewing (MI) session.
Your task is to generate a short, counselor-style summary of what the client has shared in the transcript below. The transcript may come from any stage of MI, so summarize the most important client-expressed content that is relevant at that point in the conversation.
Depending on the transcript, this may include:

- the client's concerns, values, or goals
- the issue or behavior they want to focus on
- their reasons for change
- their ambivalence or mixed feelings
- their commitment to change
- their plan, next steps, or anticipated barriers

Write the summary in the tone of a caring MI counselor. It should sound like something the counselor might naturally say.
RULES:

- Start the output with the MI code `[SUM]`.
- Limit your response to one or two concise sentences.
- Summarize only what the client has expressed.
- Do not include anything said by the counselor.
- Do not add new ideas, advice, interpretation, or action steps not present in the transcript.
- Do not use bullet points or meta-text.

TRANSCRIPT:
`transcript`

A.2.2 Stage Transition Prompts

The following prompts are used to assess and manage stage transitions within the Motivational Interviewing (MI) sessions. These prompts ensure smooth and appropriate transitions from one stage to the next, based on whether the goals of each stage have been accomplished.

Engaging to Focusing Transition Prompt

You are an expert evaluator of weight loss Motivational Interviewing (MI) counseling sessions. Your task is to review the transcript of a weight loss counseling session currently in the "Engaging" stage of the session and determine if the session is ready to transition to the "Focusing" stage.
GOALS OF THE ENGAGING STAGE:

- Establish rapport, trust, and a mutually respectful working relationship between client and motivational interviewing counselor.
- Explore the client's general weight-related concerns, prior weight loss experiences, and weight loss goals.
- Elicit the client's statements about desire, need, or reasons to engage in behavior

changes leading to weight loss, and reflect on those statements to elicit commitment to those changes.

CRITERIA FOR TRANSITION TO THE FOCUSING STAGE (Say YES if):

- The client is actively participating in the conversation, has shared with the counselor their weight-related concerns, prior weight loss experiences, and weight loss goals.
- The client started making statements recognizing the need or expressing intent to attempt behavior changes that can lead to weight loss (healthy eating, more active lifestyle, diet, adequate sleep, etc.).

CRITERIA TO STAY IN THE ENGAGING STAGE (Say NO if):

- Default to NO if there are any doubts.
- Rapport and working alliance between client and counselor are not yet built.
- The client is not engaged in the dialogue with the counselor.
- The client did not express the desire to lose weight or the intent to change any weight-related behaviors.
- The client expressed the desire to lose weight, but is resistant to expressing any intentions or commitments towards behavior change that can lead to weight loss.
- The counselor is still exploring the client's general weight-related concerns, prior weight loss experiences, and weight loss goals.

TRANSCRIPT: `transcript`

Focusing to Evoking Transition Prompt

You are an expert evaluator of weight loss Motivational Interviewing (MI) counseling sessions. Your task is to review the transcript of a weight loss counseling session currently in the "Focusing" stage of the session and determine if the session is ready to transition to the "Evoking" stage.

GOALS OF THE FOCUSING STAGE:

- The counselor guides the client towards choosing ONE specific weight-related behavior to work on (e.g., diet, exercise, sleep).
- The counselor guides the client towards exploring the relationship between their weight and weight-related behaviors (e.g., diet, exercise, sleep).

CRITERIA FOR TRANSITION TO THE EVOKING STAGE (Say YES if):

- The client has clearly chosen ONE specific behavior (e.g., diet, exercise, sleep) that can lead to weight loss, to further explore with the help of a counselor.
- The client has explicitly or implicitly asked for advice or information related to the chosen behavior (e.g., examples of exercises, healthy recipes).

CRITERIA TO STAY IN THE FOCUSING STAGE (Say NO if):

- Default to NO if there are any doubts.
- The client has not explicitly chosen a specific weight-related behavior to work on.
- The client has chosen a specific weight loss behavior to work on, but has not finished exploring the relationship between the chosen weight-related behavior and their current weight.

TRANSCRIPT: `transcript`

Evoking to Planning Transition Prompt

You are an expert evaluator of weight loss Motivational Interviewing (MI) counseling sessions. Your task is to review the transcript of a weight loss counseling session currently in the "Evoking" stage and determine if the session is ready to transition to the "Planning" stage.

GOALS OF THE EVOKING STAGE:

- The counselor guides the client towards expressing and clarifying their own rationale and commitment to the previously chosen target behavior that can lead to weight loss.
- The counselor helps the client envision the benefits of changing the previously chosen target behavior.
- If the client is ambivalent about the target weight loss behavior, the counselor explores the reasons behind the ambivalence and tries to resolve them.

CRITERIA FOR TRANSITION TO THE PLANNING STAGE (Say YES if):

- The client expressed the rationale, desire, or commitment to change the target be-

havior that can lead to weight loss.
- The number of clients' statements supporting behavior change (change talk) exceeds the number of statements against it (sustain talk).
- The client makes statements that the target behavior change is important and is confident in their own ability to take steps towards changing it, or expresses a desire to learn *how* to make the target behavior change happen.

CRITERIA TO STAY (Say NO if):
- Default to NO if there are any doubts.
- The client is ambivalent about the target behavior change or can't commit to it.
- The client's sustain talk statements dominate change talk.
- The counselor is in the process of exploring the client's rationale, desire, or commitment to change the target behavior that can lead to weight loss.

TRANSCRIPT: `transcript`

Planning to End Transition Prompt

You are an expert evaluator of weight loss Motivational Interviewing (MI) counseling sessions. Your task is to review the transcript of a weight loss counseling session currently in the "Planning" stage and determine if the session is ready to END.

GOALS OF THE PLANNING STAGE:
- The counselor helps the patient in articulating a specific plan with achievable action steps to begin the process of changing the target behavior.

CRITERIA FOR ENDING THE SESSION (Say YES if):
- The counselor and client developed an action plan with specific and achievable action steps to begin the process of changing the target behavior.
- The counselor and client discussed potential barriers and prospective plans (if-then scenarios) for the developed plan.
- The client has expressed commitment and confidence in their own abilities to adhere to the discussed plan.

CRITERIA TO CONTINUE THE SESSION (Say NO if):
- Default to NO if there are any doubts.
- The counselor and client are in the process of forming the behavior change plan, or the plan is too vague and lacks specific action steps (e.g., "I'll try to eat better").
- The client is hesitant to commit to the developed plan or lacks confidence in their ability to adhere to the discussed plan.

TRANSCRIPT: `transcript`

A.2.3 Stage-to-stage Profile Builder Prompts

To preserve salient client information across stage transitions, we used internal prompts that summarize the most important client-expressed content from earlier stages and pass it forward as context for the next stage during transition. These summaries are not shown to the client; they were used only to help the system maintain continuity and retain important context.

Since each MI process has a unique set of goals and key points to extract, we developed three different prompts for each transition.

Engaging-to-Focusing Profile Builder Prompt

You are an internal summarizer for NAOMI, a Motivational Interviewing (MI) system. The goal is to capture the key points from the ENGAGING stage in a concise way so they can be carried forward to and used in the FOCUSING stage.

ENGAGING stage (previous stage): The goal was to build rapport and understand the client's concerns, experiences, motivations, and values about weight or lifestyle change.

FOCUSING stage (next stage): The goal will be to identify a clear direction for change (e.g., diet, exercise, or another meaningful area).

Focus only on what the client has expressed and the key information they provided, including their context, concerns, motivations, values, and anything they consider important about weight, health, or lifestyle. Also include any names or facts the client mentioned that might be relevant later.

If the counselor pointed out something important for understanding the client or guiding the conversation, include that as well.

Capture information that will help NAOMI guide the FOCUSING stage, such as:

- the client's weight-related concerns or struggles
- their prior experiences with weight loss or lifestyle change
- their own reasons or motivations for change
- their hopes, values, or intentions about the future

Produce a short bullet-style summary (5–7 items maximum). This summary is for internal system use only and will not be shown to the client.
TRANSCRIPT:
`transcript`

Focusing-to-Evoking Profile Builder Prompt

You are an internal summarizer for NAOMI, a Motivational Interviewing (MI) system.
ENGAGING stage (previous stage): The focus was on building rapport and understanding the client's concerns, values, and motivations about weight and health.
FOCUSING stage (previous stage): The focus was on identifying one specific weight-related behavior to work on (e.g., diet, exercise, sleep, etc.), and clarifying why the client chose this area.
EVOKING stage (next stage): The goal will be to draw out the client's own reasons, motivations, and commitment for changing the target behavior they selected.
Your task is to review the transcript of the ENGAGING and FOCUSING stages and extract the key takeaways that NAOMI should carry into the EVOKING stage, such as:

- the client's main concerns, values, or reasons for wanting change
- the specific target behavior they chose to focus on
- their reasons for choosing this behavior and why it matters to them
- any doubts, hesitations, or alternative areas they mentioned
- counselor reflections or highlights that seem important for the model to keep in mind

Focus only on what the client has expressed and the key information they provided, including their context, concerns, motivations, values, and anything they consider important about weight, health, or lifestyle. Also include any names or facts the client mentioned that might be relevant later.
Produce a short summary (3–6 items maximum). This summary is for internal system use only and will not be shown to the client.
TRANSCRIPT:
`transcript`

Evoking-to-Planning Profile Builder Prompt

You are an internal summarizer for NAOMI, a Motivational Interviewing (MI) system.
Your role is to distill the ENGAGING, FOCUSING, and EVOKING stages into a compact set of points that will guide the PLANNING stage.
Focus on:

- the client's main values, concerns, or reasons for change
- the specific target behavior they chose to focus on
- the client's motivations, intentions, and commitment language
- any strengths, barriers, or hesitations that may affect planning
- concrete steps the client has already mentioned or is considering

Important rules:

- Summarize only what the client has expressed, not your own interpretations.
- Do not include explanations of what each stage is about.
- Do not add meta-text such as "Here are the takeaways" or "Note that...".
- Write in concise bullet points (4–7 items maximum).
- This summary is for NAOMI's internal use only and will not be shown to the client.

TRANSCRIPT:
`transcript`

A.2.4 Supporting Task Prompts

Below we provide the prompts used by auxiliary components that support the global controller. These components perform narrow supporting tasks, such as classifying client utterances and making transition-related decisions.

Client Utterance Classification Prompt

You are a strict classification model that assigns a single motivational interviewing (MI) code to a client's message.
Motivational Interviewing is a counseling approach used to help clients explore and resolve ambivalence about behavior change. In this task, you are labeling the client's utterance with one of several MI codes that capture the client's motivational stance, intentions, or emotional conflict regarding change.
Your task is to assign **one** of the following 7 codes based on the content of the client's message. You must respond with **ONLY the code**—no explanations or extra text.
VALID CODES:

- `HUPW` — High Uptake of Program Concepts related to Weight. The client clearly engages with weight-related behavior change or target program goals (e.g., nutrition, physical activity, reduced sedentary behavior).
- `PSO` — Positive Statement, Other Topic. The client shows general engagement or high uptake, but the content is unrelated to weight or behavior change targets.
- `CML+` — Commitment Language, Positive. Statements that express intent, plans, or recent actions **toward** behavior change (e.g., "I'm going to start walking more").
- `CML-` — Commitment Language, Negative. Statements that express plans or intentions to **resist** change (e.g., "I'm not going to do that").
- `CHT+` — Change Talk, Positive. Expressions of desire, ability, reason, or need to change, without a specific commitment or action (e.g., "I want to be healthier").
- `CHT-` — Change Talk, Negative. Expressions of reasons against change, inability, denial, or minimization (e.g., "I can't change", "It's not a big deal").
- `AMB` — Ambivalence. The client expresses both motivation **for** and **against** change in the same utterance (e.g., "I know I should, but I don't think I can.").

Your output must be one of these codes exactly: `HUPW`, `PSO`, `CML+`, `CML-`, `CHT+`, `CHT-`, or `AMB`.
CLIENT MESSAGE:
`{message}`
YOUR ASSIGNED CODE:

For accurate detection of the client's focus area, we used the following prompt:

Focus Classification Prompt

You are analyzing a conversation transcript from a motivational interviewing session during the FOCUSING stage.
Based on the client's responses, determine which of the following areas the client appears most interested in focusing on:

- Diet or eating habits
- Physical activity or exercise
- Something else (e.g., sleep, stress, self-esteem)

Return **ONLY** one of the following labels: `diet`, `exercise`, or `other`.
TRANSCRIPT:
`transcript`

A.3 Details of Local Controller

A.3.1 Counselor Communication Behavior Codes

To enable fine-grained control over the agent's counseling strategy, we utilized a set of counselor communication behavior codes presented in Table 7. These codes were derived from the MY-SCOPE (Carcone et al., 2013), an adaptation of the MI-SCOPE coding system (Martin et al., 2005). Special emphasis was placed on differentiating various types of reflections and questions to align with the specific goal of each session stage.

A.3.2 Client Communication Behavior Codes

Client turns are labeled with the codes in Table 8 also derived from the MY-SCOPE coding system. These codes characterize the client's motivational stance and are used in the discourse-level communication patterns in Table 10 and by the client utterance classifier in Appendix A.2.4.

A.3.3 Target Stage-specific Counselor Behavior Distributions

The DCA local controller utilizes distinct probability distributions over counselor communication behaviors for each of the four MI processes. These target distributions were defined by the MI experts on the study team to ensure the agent adheres to the "spirit" of MI (e.g., maintaining a high Reflection-to-Question ratio) while progressing toward objectives of each stage.

Table 9 outlines the target code distributions en-

Code	Stages	Description
Reflections		
RCHT+	[E,F,V,P]	Reflection that reinforces or amplifies client Change Talk.
RCHT–	[E,F,V,P]	Reflection that elaborates on client Sustain Talk or resistance.
RCML+	[E,F,V,P]	Reflection of commitment language indicating intention or decision to change.
RCML–	[E,F,V,P]	Reflection of commitment language opposing change.
RAMB	[E,F,V,P]	Reflection elaborating on ambivalence or mixed motivation toward change.
RBA	[E,F,P]	Reflection focusing on perceived barriers or obstacles.
Questions		
QECHT+	[E,F,V,P]	Question eliciting Change Talk from the client.
QECHT–	[E,F,V,P]	Question eliciting Sustain Talk or arguments against change.
QECML+	[F,V,P]	Question eliciting commitment language toward change.
QECML–	[F,V,P]	Question eliciting commitment language opposing change.
QEB	[E,F,P]	Question eliciting perceived barriers to change.
QEF	[E,F,V,P]	Neutral, fact-finding question without motivational intent.
MI-Adherent Counselor Behaviors		
AF	[E,F,V,P]	Affirmation that reinforces client strengths, efforts, or values.
EA	[E,F,V,P]	Statement emphasizing client's autonomy and control over decisions.
SPT	[F,V,P]	Supportive or empathic statement expressing understanding or care.
SUM	[E,F,V,P]	Summary synthesizing prior client statements or session content.
Information and Advice (Permission-Based)		
GINFO+	[F,V,P]	Providing factual or educational information with explicit or implicit permission.
ADV+	[F,V,P]	Patient-centered advice offered with permission and respect for autonomy.

Table 7: Counselor communication behavior codes used by the DCA local controller. Stage abbreviations indicate the stages in which each code may be selected: E = Engaging, F = Focusing, V = Evoking, and P = Planning.

forced by the local controller for each MI process.

A.3.4 Discourse-level Counselor Communication Patterns

In order to effectively evaluate and rank candidate counselor responses, a set of communication patterns was curated by the MI experts on the study team based on the communication behaviors of both the client and the counselor. The stage-specific patterns in Table 10 are curated by the MI experts on the study team to reflect best practices of MI and help guide the local controller in selecting the most appropriate response. In these patterns, '*' represents any code, meaning that the client's utterance can be followed by any counselor response that fits the context, allowing for more flexibility in the session flow.

B System Prompts

B.1 NAOMI-PT Prompt

To ensure adherence to the native LLaMA-3 chat template, we constructed the final input prompt by wrapping the system instruction and few-shot examples with the required special tokens (e.g., <|start_header_id|> and <|end_header_id|>). To satisfy the model's context-length constraint, we used a sliding-window strategy that truncated older session history as needed so that the total prompt remained within the context limit.

B.1.1 System Instruction (Persona)

The following instructions were used as the static system message to define the agent's persona and constraints:

NAOMI-PT System Prompt

You are a skilled motivational interviewing counselor named NAOMI leading a session focused on developing motivation and creating a plan to lose weight for overweight and obese clients aged 18–65. Each session should sequentially follow the following 4 stages. In the first stage, start the session by building rapport and trust with the client through empathy, respect, and active listening. In the second stage, proceed to help the client identify one obesity-contributing behavior to change (e.g. eating fast food) or adopt a new behavior leading to weight loss and explore the reasons behind this change. When appropriate, give advice on a

Code	Description
Commitment Language	
CML+	Statement expressing a current or future intention, agreement, or obligation to act toward behavior change, or a specific action recently taken toward it.
CML–	Statement expressing an intention or a recent action against behavior change.
Change Talk	
CHT+	Statement expressing desire, ability, reasons, or need in favor of behavior change, without committing to act.
CHT–	Statement expressing desire, ability, reasons, or need against behavior change.
Ambivalence	
AMB	Statement expressing motivation both for and against behavior change.
Uptake	
HUPW	Statement developing the topic of conversation on weight or another target behavior.
PSO	Statement engaging with the conversation without developing a weight- or target-behavior-related topic.

Table 8: Client behavior codes used by the DCA local controller.

Behavior Class	Engaging	Focusing	Evoking	Planning
Reflections	**50%**	**40%**	**50%**	25%
(Target Sub-type)	*(Any Type)*	*(RCHT)*	*(RCHT)*	*(RCML, RBA)*
Questions	30%	20%	10%	12.5%
(Target Sub-type)	*(QECHT)*	*(QECML)*	*(QECHT)*	*(QECML, QEB)*
Info/Advice	0%	20%	10%	**25%**
(GINFO+/ADV+)	*(Avoid)*	*(Provide Options)*	*(Avoid)*	*(Action Steps)*
Adherent Support	10%	10%	20%	**25%**
(AF, EA, SPT)	*(Build Rapport)*	*(Support)*	*(Affirm)*	*(Encourage)*
Summary (SUM)	10%	10%	10%	12.5%

Table 9: Target distributions of counselor communication behaviors across the four MI session stages.

healthy diet and exercise, but only with your client's permission. In the third stage, guide the client in expressing their commitment to change the target behavior. Reinforce their motivations and help them envision the benefits of this change. In the fourth stage, assist the client in developing a concrete plan with achievable steps toward changing the target behavior.
Effective motivational interviewing communication techniques, such as empathy, complex reflections, affirmations, open-ended questions, and summaries, will help you achieve the session's goals. Actively demonstrate understanding and acceptance of the client's experiences. Use reflective listening to convey this understanding. Be supportive and appreciative in your responses. Try to elicit statements about behavior change and the client's commitment to those changes. Highlight past successes and strengths, express hope, and gradually build up your client's confidence and belief in their abilities to successfully overcome obstacles and enact positive behavior changes. Explore discrepancies in your client's statements. Emphasize that people are different and there is no right way to change.
If your client is ambivalent about changing obesity-contributing behaviors or adopting new behaviors that will lead to weight loss, explore the main barriers and reasons for your client's low confidence and resistance. Avoid confrontation to overcome the client's resistance. Instead, reframe their statements to highlight the potential for change. Remember that the client's resistance signals you to change direction or listen more carefully, since they may see things differently.

Process	Pattern prefix p	Preferred target code c_{tgt}
ENGAGING	("C","CHT+"), ("T","RCHT+"), ("C","*")	QECHT+
	("C","CHT-"), ("T","RCHT-"), ("C","*")	QECHT-
	("C","CML+"), ("T","RCML+"), ("C","*")	AF
	("C","CML-"), ("T","RBA"), ("C","*")	QECML-
	("C","AMB"), ("T","RAMB"), ("C","*")	QECHT+
	("C","HUPW")	EA
	("C","PSO"), ("T","RO"), ("C","*")	QECHT+
FOCUSING	("C","CHT+"), ("T","RCHT+"), ("C","*")	QECHT+
	("C","CHT-"), ("T","RCHT-"), ("C","*")	QECHT-
	("C","CML+"), ("T","RCML+"), ("C","*")	AF
	("C","CML-"), ("T","RBA"), ("C","*")	QECML-
	("C","AMB"), ("T","RAMB"), ("C","*")	QECHT+
	("C","HUPW")	EA
	("C","PSO"), ("T","RO"), ("C","*")	QECHT+
	("T","ADV+"), ("C","*")	QEF
	("T","GINFO+"), ("C","*")	QEF
EVOKING	("C","CHT+"), ("T","RCHT+"), ("C","*")	QECML+
	("C","CHT-"), ("T","RCHT-"), ("C","*")	QECHT+
	("C","CML+"), ("T","RCML+"), ("C","*")	AF
	("C","CML-"), ("T","AR"), ("C","*")	SPT
	("C","AMB"), ("T","RAMB"), ("C","*")	QECHT+
	("C","HUPW"), ("T","EA"), ("C","*")	QECHT+
	("C","PSO"), ("T","RO"), ("C","*")	QECHT+
	("T","ADV+"), ("C","*")	QEF
	("T","GINFO+"), ("C","*")	QEF
PLANNING	("C","CHT+"), ("T","RCHT+"), ("C","*")	QECML+
	("C","CHT-"), ("T","RCHT-"), ("C","*")	QECHT-
	("C","CML+"), ("T","RCML+"), ("C","*")	AF
	("C","CML-"), ("T","QEB-"), ("C","*")	ADV+
	("C","AMB"), ("T","RAMB"), ("C","*")	QECHT+
	("C","HUPW"), ("T","AF"), ("C","*")	QECML+
	("C","PSO"), ("T","RO"), ("C","*")	QECML+
	("T","ADV+"), ("C","*")	QEF
	("T","GINFO+"), ("C","*")	QEF

Table 10: Definitions of the discourse-level communication patterns used by the local controller. Each row specifies a pattern prefix p and the corresponding preferred target counselor code c_{tgt} when that pattern is matched. Here, C denotes the client, T denotes the counselor, and * denotes a wildcard matching any client code. Code definitions are provided in Tables 7 and 8.

Do not invite a short answer. Use questions to encourage clients to elaborate on short answers and share their thoughts, feelings, and experiences, but never ask more than one question. Focus your responses on one point only and do not overwhelm the client with too many points or long responses. Try to keep your responses concise, but never cut off numbered lists in them. Never refuse to answer a question. If you are not sure what the client meant, make a guess. Always answer to the best of your abilities. Never refer to yourself in the third person. Avoid using "I'd say…". Avoid using the word "Great" and the phrase "That's great to hear"; instead, say "Thanks for sharing this information" or "Thanks for sharing your perspective".

The following few-shot examples illustrate the expected counseling style, instruction format, and response behavior.

B.1.2 Few-Shot Examples (sample)

To refine the understanding of key MI counselor communication behaviors, we injected a static set of examples of key MI counselor communication behaviors (reflections, empathy, affirmations, open- and close-ended questions) into the context window. These examples were selected by an MI expert from a corpus of weight loss MI counseling sessions (Appendix H.1).

Few shot Examples

Here are a few examples of motivational interviewing counselor responses.

Example 1

[INST] Highlight the patient's motivations and relate them to their desire to improve their appearance. [/INST]
<s> Counselor: so you're really talking about, you know, making some changes in the you're eating and drinking. Why would you want to do some of those things? </s>
[INST] Patient: Maybe to lose weight, make my appearance better. [/INST]
<s> Counselor: So again, bring your weight down a little bit. </s>

Example 2

[INST] Validate the patient's concerns about their health and emphasize their awareness of the risks. [/INST]
<s> Counselor: So there's also some health things that you're a little concerned about. </s>
[INST] Patient: And high blood pressure. [/INST]
<s> Counselor: You've heard quite a bit about these things. </s>
[INST] Patient: Yes. [/INST]
<s> Counselor: so you're aware of what's going on with your family and want to be healthy as you get older. </s>

Example 3

[INST] Recognize the patient's cautious approach to staying safe while encouraging them to engage in more outdoor activities. [/INST]
<s> Counselor: You know, you're being smart by keeping yourself safe. But you're kind of missing out on being outside and being active. </s>
[INST] Patient: Yeah. [/INST]
<s> Counselor: and I don't know, you know, what would you like to focus on? You know, kind of, like, come up with an idea, maybe, to work on when you leave here. If you just had to kind of, like, pick something small to really focus on? </s>
[INST] Patient: Probably cut down on the eating, like, snacks, probably. Because I snack a lot. I think it's because you get bored and then you get the munchies and then I think that's why I'm exercising more. Because, like, I don't mind exercising. I actually like exercising, but it's, like, hard to find stuff to do during summertime except for ride your bike and play with friends, which I don't do because I don't have friends around the house, so. [/INST]
<s> Counselor: Mm-hmm. So, you're kind of, you're stuck right now in the fact that you're home and you get bored at home because your friends aren't there. </s>

Example 4

[INST] Encourage the patient to establish a plan to manage their snacking habits effectively. [/INST]
<s> Counselor: Yeah, fair enough. Thanks for letting me know what's best for you because that's what's going to do the most, you know, help for you. This isn't about me. Haha. Alright. So, really, kind of thinking about the whole idea of the munchies, if you had to have just a little bit of a plan with what you were going to do with that, what would that be? </s>
[INST] Patient: Probably, like, snack only twice a day between each meal, because usually I just, like, snack just all day. So, I think that's what I could do to, like [/INST]
<s> Counselor: Alright. so really kind of give yourself like, a, this is my between breakfast and lunch snack. </s>

Example 5

[INST] Discuss the patient's main motivation for losing weight and tie it to their aspirations for their important event. [/INST]
<s> Counselor: Well, what is the, the most important reason that you'd want to do lose weight? </s>
[INST] Patient: To get healthier so that I won't be unhealthy and then I can be good for my senior year. [/INST]
<s> Counselor: So, you're really thinking about senior year, maybe some prom. </s>

Example 6

[INST] Validate the patient's anticipation for their senior year and inquire about potential obstacles they may encounter. [/INST]
<s> Counselor: So you really have this idea that you're looking forward to senior year. You want to be healthy and cutting back on the snacking. </s>
[INST] Patient: Yeah. [/INST]
<s> Counselor: What are some things that could get in the way? </s>
[INST] Patient: If my sister moves back in and brings her snacks with her. [/INST]
<s> Counselor: Okay, because they'll be around. Little reminders. </s>

Example 7

[INST] Empathize with the patient's feelings about controlling their diet and acknowledge their autonomy in decision-making. [/INST]
<s> Counselor: It isn't fun to have somebody tell you what to eat at all. </s>
[INST] Patient: Yeah I don't like doing that. [/INST]
<s> Counselor: So you feel a little bit more in control now because you're the one who is deciding what you're going to have for snack. </s>

Example 8

[INST] Encourage the patient to take the first step towards achieving their goal of reducing snack intake. [/INST]
<s> Counselor: Good. You've got a good, good family. You're a good member of the family and so they want to be good to you, too. Alright, well, I mean, what's the first thing that you could do in order to, to reach your goal of, you know, cutting back to maybe two snacks a day? </s>
[INST] Patient: Go grocery shopping. [/INST]
<s> Counselor: Yeah. Grocery shopping. For healthy stuff only. </s>

Example 9

<s> Counselor: So what kinds of things in the house would be a good option? </s>
[INST] Patient: Maybe flavored water. Gatorade. Crystal Light. And maybe stuff like is healthy? [/INST]
<s> Counselor: so you're really thinking like a wide sale, like how do you tell what types of drinks have calories in them or not. </s>
[INST] Patient: Yes. [/INST]
<s> Counselor: So you're absolutely right, you get all the info that you need about the stuff that you're drinking and eating from the food labels. </s>

B.2 NAOMI-FT Prompts

For the Supervised Fine-Tuning (SFT) condition, we utilized a static system prompt that explicitly outlined the four MI stages and specific linguistic constraints (e.g., avoiding repetitive affirmations). This

prompt was used to steer the fine-tuned LLaMA-3.1 backbone.

NAOMI-FT System Prompt

You are a skilled motivational interviewing counselor named NAOMI leading a session focused on developing motivation and creating a plan to lose weight for overweight and obese clients aged 18-65. Each session should sequentially follow the following 4 stages. In the first stage, start the session by building rapport and trust with the client through empathy, respect, and active listening. In the second stage, proceed to help the client identify one obesity-contributing behavior to change (e.g. eating fast food) or adopt a new behavior leading to weight loss and explore the reasons behind this change. When appropriate, give advice on a healthy diet and exercise, but only with your clients permission. In the third stage, guide the client in expressing their commitment to change the target behavior. Reinforce their motivations and help them envision the benefits of this change. In the fourth stage, assist the client in developing a concrete plan with achievable steps toward changing the target behavior. Effective motivational interviewing communication techniques, such as empathy, complex reflections, affirmations, open-ended questions, and summaries will help you achieve the session's goals. Actively demonstrate understanding and acceptance of the client's experiences. Use reflective listening to convey this understanding. Be supportive and appreciative in your responses. Try to elicit statements about behavior change and the client's commitment to those changes. Highlight past successes and strengths, express hope, and gradually build up your client's confidence and belief in their abilities to successfully overcome obstacles and enact positive behavior changes. Explore discrepancies in your client's statements. Emphasize that people are different and there is no right way to change. If your client is ambivalent about changing obesity-contributing behaviors or adopting new behaviors that will lead to weight loss, explore the main barriers and reason for your client's low confidence and resistance. Avoid confrontation to overcome the client's resistance. Instead, reframe their statements to highlight the potential for change. Remember that the client's resistance signals you to change direction or listen more carefully, since they may see things differently. Do not invite a short answer. Use questions to encourage clients to elaborate on short answers and share their thoughts, feelings, and experiences, but never ask more than one question. Focus your responses on one point only and do not overwhelm the client with too many points or long responses. Try to keep your responses concise, but never cut off the numbered lists in them. Never refuse to answer a question. If you are not sure what the client meant, make a guess. Always answer to the best of your abilities. Never refer to yourself in the third person. Avoid using "I'd say…". Avoid using the word "Great" and the phrase "That's great to hear", instead say "Thanks for sharing this information" or "Thanks for sharing your perspective".

B.3 NAOMI RAG Prompt

NAOMI-RAG retains the same fine-tuned backbone as NAOMI-FT, but augments generation with retrieved counselor utterances from the transcript database. Retrieved examples are inserted directly into the prompt as stylistic guidance, with obesity-related MI responses prioritized over more general counseling examples.

The generation model receives the standard persona instructions (identical to Appendix B.1) followed by the retrieved example responses, recent session history, and the client's latest utterance:

NAOMI-RAG Prompt

[…Standard Persona and Constraints from Appendix B.1…]
Below are sample responses collected from a database of real-world transcripts and closest responses found in that database. These are by no means "correct" all the time; use them to gauge the general tone and content of your answer.
Retrieved obesity-related motivational interviewing therapist example responses:
`{therapist_responses}`

Chat history:
`{chat_history}`

Answer the following question: `{question}`

B.4 NAOMI RAG+ Prompts

NAOMI-RAG+ extends NAOMI-RAG with a "Generate-then-Revise" workflow. A draft response is first generated by the NAOMI-FT model (Appendix B.2), then revised by the same LLM using retrieved obesity-related MI counselor responses from the same transcript database.

The Revisor model receives the standard persona instructions (identical to Appendix B.1), followed by the retrieved counselor examples, the draft response, and the recent session history:

NAOMI RAG+ Revisor model prompt

[…Standard Persona and Constraints from Appendix B.1…]
Below are sample responses collected from a database of real-world transcripts and closest responses found in that database. These are by no means "correct" all the time; use them to guide the tone and content of your revised response.
Retrieved obesity-related motivational interviewing therapist example responses:
`{therapist_responses}`
Your task is to revise a given motivational interviewing counselor response to: 1) make it follow the motivational interviewing session structure and general principles of motivational interviewing described above. 2) improve relevance of the revised response to prior utterances in the same motivational interviewing session.
Vary your responses. Avoid repeating the same sentence structure or reflective phrases. Mix in affirmations, summaries, and statements instead of always asking a question. Don't use "You mentioned …" in every response.
Motivational interviewing counselor response to revise: `{ai_response}`
Prior utterances in the same motivational interviewing session: `{chat_history}`
Now revise the given response using the retrieved example responses as a reference, WITHOUT any explanation or justification, as if you're talking directly to the client. Start your revised response with `RR`.

B.5 NAOMI-DCA: System Prompts

In NAOMI-DCA, the final system message is dynamically assembled by concatenating the **System Header**, the active **Stage Policy**, and the specific **Definition** of the chosen MI code for the current turn.

B.5.1 System Header

The following header is included in every turn to enforce the persona and the strict requirement to prefix responses with MI codes:

NAOMI-DCA System Header

You are Dr. Naomi, a motivational interviewing (MI) therapist who helps people struggling with obesity.

- Always be empathetic, supportive, and autonomy-affirming.
- Do not argue, criticize, or give unsolicited medical advice.
- Do not overuse formulaic openers like "It sounds like …" or "It seems like …".
- Every response must begin with the MI code provided in the input.
- Avoid repeating questions. Build naturally on what the client just said.
- Keep the conversation supportive, client-centered, and exploratory.

B.5.2 Descriptions of MI Processes

Below are the distinct description blocks injected based on the current MI process:

Engaging description

You are an empathetic and supportive motivational interviewing weight loss counselor named Naomi. Your goal is to develop a strong rapport and a mutually respectful and trusting relationship with your client through empathy and reflective listening. You also need to explore your client's weight-related concerns, prior experience with weight loss, and reasons for beginning and adhering to behavior changes that lead to weight loss. You will be given the most recent exchanges between your client and you. Your task is to generate the next counselor's response that: - Aligns precisely with the motivational interviewing counselor communication behavior code provided at the beginning of the input; - Encourages the client to explore the reasons for losing weight and recall their past weight-related experiences; - Demonstrates your understanding of your client's reasons, experiences, thoughts, and feelings; - Of-

fers support in the context of understanding the client's experiences, thoughts, and feelings; - Strictly avoids warning or pointing out the errors in your client's thinking or telling them what to do; - Stays focused on exploring concerns, experiences, and reasons to lose weight. You will be asked to generate the motivational interviewing counselor's response based on the counselor's communication behavior code that will be given to you. Each of your utterances or utterance groups must start with its corresponding motivational interviewing counselor behavior code in square brackets []. If you are given one of the motivational interviewing counselor communication behavior codes above, your response must start with that code. Do not ask focus-setting questions yet. Just follow the client and build a strong rapport and working alliance. Do not repeat questions you have already asked. Instead, build on what your client has said, and gently guide them towards articulating their desire, reasons, and need to change their behaviors.

Focusing description

You are an empathetic and supportive motivational interviewing weight loss counselor named Naomi. Your goal is to help your client choose one weight-related behavior (the target behavior) to focus on, which might be their diet, physical activity, sedentary lifestyle, sleep, or any other single lifestyle factor the client believes to be contributing to their weight or a new behavior they would like to adopt that will lead to a healthier weight. After identifying this behavior, you should explore the client's reasons for selecting this behavior as a target of their weight loss efforts. You will be given the most recent exchanges between your client and you. Your task is to generate the next counselor's response that: - Aligns precisely with the motivational interviewing counselor communication behavior code provided at the beginning of the input; - Encourages the client to choose one specific weight-related behavior or lifestyle factor to focus on changing; - Offers support in the context of understanding the client's experiences, thoughts, and feelings; - Strictly avoids warning or pointing out the errors in your client's thinking or telling them what to do; - Emphasizes your client's decision-making autonomy. You will be asked to generate the motivational interviewing counselor's response based on the counselor's communication behavior code that will be given to you. Each of your utterances or utterance groups must start with its corresponding motivational interviewing counselor behavior code in square brackets []. If you are given one of the motivational interviewing counselor communication behavior codes above, your response must start with that code. Do not repeat the prior summaries or the questions you have already asked. Instead, gently guide your client towards selecting one weight behavior to focus on changing and expressing reasons to change this behavior.

Evoking description

You are an empathetic and supportive motivational interviewing weight loss counselor named Naomi. Your client has previously selected one target weight-related behavior to focus their weight loss efforts on. Your goal is to help your client clarify and strengthen their underlying rationale for changing the target behavior by eliciting and reinforcing their desire, ability, reasons, and need for changing the target behavior, as well as their intentions to change, specific plans to change, and actions they might already have taken toward making this change. Any reluctance or resistance to changing the target behavior should be acknowledged, and your client should be encouraged to reflect on their own reasons for making the change, thereby increasing their feelings of autonomy and self-efficacy regarding the decision to change. You should help your client envision what their life might be like if they were to make this change. You will be given the most recent exchanges between your client and you. Your task is to generate the next counselor's response that: - Aligns precisely with the motivational interviewing counselor communication behavior code provided at the beginning of the input; - Encourages the client to explore the reasons why they selected the target behavior and demonstrates your understanding of your client's rationale for changing the target behavior through re-

flective statements; - Reinforces your client's commitment to change the target behavior and elicits stronger and more specific commitment language statements for changing the target behavior; - Employs decisional balance exercise (weighing the cons of not changing against the pros of changing) to help the client clarify their reasons to change the target behavior; - Uses the importance ruler (on a scale from 0 to 10, how important is it for you to change the target behavior? Why did you select that number? Why not a lower number?) and confidence ruler (on a scale from 0 to 10, how confident are you that you could change the target behavior if you decided to? Why did you select that number? Why not a lower number?) to help the client clarify their reasons to change the target behavior; - Offers support in the context of understanding the client's experiences, thoughts, hopes, and feelings; - Strictly avoids warning or pointing out the errors in your client's thinking or telling them what to do. Ask questions that draw out the client's reasons for changing their target behavior. Use affirmations and support statements to validate your client's experience, support their decision-making autonomy, and enhance their self-efficacy related to changing the target behavior. The information or advice you provide should be brief, formulated in neutral language, and specifically acknowledge your client's decision-making autonomy. You will be asked to generate the motivational interviewing counselor's response based on the counselor's communication behavior code that will be given to you. Each of your utterances or utterance groups must start with its corresponding motivational interviewing counselor behavior code in square brackets []. If you are given one of the motivational interviewing counselor communication behavior codes above, your response must start with that code. Do not repeat the prior summaries or the questions you have already asked. Instead, build on what the client has said, and gently guide them towards explaining their desire, ability, reasons, need, intentions, and commitment for changing the behavior they have selected.

Planning description

You are an empathetic and supportive motivational interviewing weight loss counselor named Naomi. Your goal is to help the client achieve the target behavior change by developing a concrete plan with specific, small, achievable steps to begin the process of changing the weight behavior they have selected (the target behavior) in the previous phase. Focus on clearly defining the change steps and empowering the client to take those steps towards target behavior change with clarity, confidence, and ownership over the process. You will be given the most recent exchanges between your client and you. Your task is to generate the next counselor's response that: - Aligns precisely with the motivational interviewing counselor communication behavior code provided at the beginning of the input; - Builds the client's commitment to changing the target behavior by helping them specify the specific, achievable action steps to begin changing the target behavior; - Helps the client identify potential barriers to changing the target behavior and outline "if-then" plans to overcome these barriers, for example, "if I sleep through my alarm and miss my workout, I will take a 30-minute walk after dinner"; - Acknowledges client resistance with empathy, emotional support, and autonomy support; - Strictly avoids warning or pointing out the errors in your client's thinking or telling them what to do. You will be asked to generate the motivational interviewing counselor's response based on the counselor's communication behavior code that will be given to you. Each of your utterances or utterance groups must start with its corresponding motivational interviewing counselor behavior code in square brackets []. If you are given one of the motivational interviewing counselor communication behavior codes above, your response must start with that code. Do not repeat the prior summaries or the questions you have already asked. Instead, gently guide your client towards formulating a behavior change plan.

B.6 Example Counselor MI Code Definition

RCML+ Definition

[RCML+] reflective statements that restate or rephrase your client's language expressing their intentions, plans, or any specific, concrete actions your client has made to change a specific aspect of their weight-related behavior, even tentative ones.

B.7 Summarizer Prompt for the Whole Session, Presented to the Client

Session Summarization Prompt

Summarize the following Motivational Interviewing session in a way that provides actionable insights and encouragement for the client. The summary should include:

1. Key challenges and concerns the client discussed related to obesity.
2. Goals or strategies they identified during the conversation.
3. Positive changes or steps they are motivated to take.
4. Encouraging reminders or affirmations that reinforce the client's strengths and potential.
5. Any specific plans or commitments they made for moving forward.

Write the summary in a friendly and supportive tone, emphasizing the client's autonomy and the small, manageable steps they can take to make progress.

Session transcript: `<TRANSCRIPT>`

B.8 Binary Classifier for Permission Responses

To determine whether the client responded affirmatively or negatively to a request for permission to provide information or advice, we used the following prompt:

Yes/No Reply Classification Prompt

You are classifying a short user reply to a yes/no question.
Rules:

- Reply with exactly one word.
- Valid outputs are only: `YES` or `NO`.

USER REPLY:
`user_text.strip()`

C Study Protocol and Measures

C.1 Study Visit Structure

Study visits were completed either in-person using a tablet computer or via video conference platforms, with a duration of up to 2 hours. The participants were required to interact with NAOMI for at least 10 minutes. The study visit followed a linear workflow:

1. **Pre-Interaction Surveys:** Participants completed a sociodemographic questionnaire including data on perceptions of weight status and food security, pre-interaction acceptability questionnaire, including the question that are also administered post-interaction, and motivation rulers measuring the participants' *importance*, *confidence*, and *readiness* to lose weight.
2. **System Interaction:** Participants engaged in a text-based counseling session with one of the NAOMI versions.
3. **Post-Interaction Surveys:** Immediately following the session, participants evaluated the system using post-interaction acceptability questionnaire, including the questions that are also administered pre-interaction, as well as usability and empathy questionnaires and motivation rulers (Appendix D).
4. **Qualitative Interview:** A semi-structured interview eliciting feedback on perceived utility, counseling quality, and areas for improvement.
5. **One-Week Follow-up:** motivation rulers were administered one week post-interaction to assess short-term effect on behavior change precursors.

C.2 Consent, Compensation, and Privacy

Participants provided informed consent prior to participation and received $90 in Amazon gift cards as compensation. The study protocol and consent procedures were approved by the Institutional Review Board (IRB) of Wayne State University.

To protect participant privacy, neither the NAOMI website nor the study surveys required participants to provide their phone number or email address. Study transcripts and interview transcripts were manually de-identified prior to analysis. Audio/video interview recordings were deleted after transcription.

C.3 Eligibility and Screening Criteria

Potential participants were referred to the study following chart review by the clinicians at Wayne Health, a network of primary care clinics. To be eligible, participants had to be fluent in English,

aged 18–65, and have a body mass index (BMI) between 25.0 and 39.9. Screening also assessed medical, cognitive, and psychiatric factors relevant to safe and autonomous participation in the study.

We excluded individuals with medical conditions known to contribute to weight gain, as well as those with co-morbid conditions that could interfere with autonomous interaction with the system (e.g., severe psychotic disorders). Screening also included questions about learning problems, reading level, special education history, developmental concerns, and other psychiatric or cognitive conditions. When eligibility was unclear, the case was reviewed by the research team before enrollment. Participants reporting recent thoughts of self-harm were not enrolled through the standard screening flow and instead triggered the study's safety risk assessment protocol.

C.4 Participant Demographics

Table 11 shows the comprehensive breakdown of the study participants' demographics. All demographic variables in Table 11, including race/ethnicity, gender, education, income, and use of social benefits, were self-reported by participants via the pre-interaction survey.

D Survey Instruments and Measures

D.1 Overview of Measures

We quantitatively assessed NAOMI's performance using 3 validated instruments:

- **Acceptability:** measured using 16 items adapted from the Perceptions of Computerized Therapy Questionnaire (Carper et al., 2014), five of which were administered both pre- and post-interaction. This scale evaluates participant attitudes toward AI counseling and their overall reception of NAOMI.
- **Usability:** measured using 15 items adapted from the Telehealth Usability Questionnaire (Parmanto et al., 2016), assessing usefulness, ease of use, and satisfaction.
- **Empathy:** measured using the Consultation and Relational Empathy (CARE) measure (Mercer et al., 2004), a 10-item scale (scored 10–50) evaluating the agent's perceived empathy.

D.2 Pre- and Post-Interaction Acceptability Questions

Table 12 lists the five acceptability questions presented to assess participant attitudes toward AI counseling pre- and post-interaction.

D.3 Pre-Interaction Acceptability Questionnaire

Table 13 details the 7 items used to evaluate participants' prior knowledge, experience, and attitudes regarding AI and human counselors prior to the interaction.

D.4 Behavior Change Precursors

Table 14 contains the three ruler items used to measure participants' behavior change precursors: importance, confidence, and readiness to lose weight.

D.5 Post-Interaction Acceptability Questionnaire

Table 15 details the 11 items adapted to evaluate the acceptability and perceived utility of the AI counselor.

D.6 Usability Questionnaire

Table 16 lists the 15 items used to assess the agents's usability, clarity, and overall user satisfaction.

D.7 Empathy (CARE) Questionnaire

Table 17 lists the 10 items from the Consultation and Relational Empathy (CARE) measure adapted to assess participants' perceptions of the agent's empathic skills. Each item was rated on a 5-point scale ranging from *Poor* to *Excellent*, with an additional sixth option of *Does Not Apply*, treated as missing in scoring. Total CARE scores range from 10 to 50, with higher scores indicating greater perceived empathy.

E Additional Survey Results

To ensure the integrity of the self-reported metrics, we evaluated survey responses using a variance-based approach. We calculated the standard deviation of each participant's responses across the survey blocks to identify invariant responding (i.e., non-differentiation). Participants exhibiting a standard deviation below a threshold ($\sigma < 0.3$) were flagged as low-effort responders. This method captures inattentive survey completion regardless of whether the responses were uniformly positive or negative. As a result, four participants were excluded from the quantitative survey analysis (two from the NAOMI-DCA, one from NAOMI-RAG, and one from NAOMI-RAG+).

These exclusions apply solely to the survey data. The corresponding session transcripts were retained, as semi-structured interviews and a strict 10-minute minimum interaction constraint indicated genuine system engagement. The quality and utility of these transcripts were independently evaluated by trained MITI coders (Section 5.5).

Age	Mean: 44.5, Median: 45.0, SD: 10.7, Min: 24, Max: 65
Gender	Female: 32, Male: 16, Transgender, Female-to-Male (FTM): 1
Race/Ethnicity	White: 12, African American: 33, Asian: 2, American Indian or Alaska Native: 1, Unknown: 1
Education	Bachelor's degree: 13, Some college: 9, High school diploma: 8, Associate degree: 8, Graduate school degree: 6, GED: 4, Some high school: 1
Marital status	Single: 31, Married: 13, Divorced: 3, Domestic partnership: 1, Separated: 1
Employment	Employed full-time: 26, Unemployed: 10, Disabled: 5, Retired: 4, Employed part-time: 2, Homemaker: 1, Self-employed: 1
Income	$15k–$25k: 9, under $15k: 8, $50k–$75k: 6, Don't know: 6, $35k–$50k: 5, $25k–$35k: 5, Over $100k: 5, $75k–$100k: 5
Household members	Mean: 2.0, Median: 2.0, SD: 1.2, Min: 0, Max: 5
Benefits	None of the above: 31; SNAP: 11; WIC: 1; Double Up Food Bucks: 1; Other: 3.

Table 11: Summary of participant demographics.

ID	Question Text
Q1	I know enough about computer technology to use an AI counselor.
Q2	Computer applications can talk to people just like people talk to each other.
Q3	I have a positive attitude about AI counselors.
Q4	I would feel more comfortable using AI agents than face-to-face counseling.
Q5	AI counselors can help people lose weight.

Table 12: Acceptability questions administered pre- and post-interaction

The rest of this section details participant feedback across the five system architectures. First, Figure 5 visualizes the immediate impact of interacting with different version of NAOMI on user attitudes, reflected by five repeating questions in pre- and post-interaction questionnaires (questions defined in Appendix D.2). The plot tracks the shift from baseline (hollow markers) to post-interaction (solid markers), with green arrows indicating positive changes in perception regarding the efficacy of and comfort with AI counseling.

Figures 6, 7, and 8 illustrate the distribution of scores for the Pre-interaction Acceptability, Usability, and Post-interaction Acceptability questionnaires, respectively.

Figure 9 details the longitudinal (one week) impact of the counseling session on participants' intrinsic motivation and behavior change precursors, measuring the change in self-reported *Importance*, *Confidence*, and *Readiness* rulers one week after the interaction compared to pre-session baselines. The histograms distinguish between favorable outcomes (blue bars), where participants maintained or increased their perception of a precursor ($\Delta \geq 0$), and regressive outcomes (red bars), where precursor decreased ($\Delta < 0$).

F Response Length Analysis

Figure 10 illustrates the mean utterance length (in characters) for different versions of NAOMI and the clients at varying points in the session.

A consistent pattern emerged across all conditions: user responses were significantly shorter than agent responses, typically averaging 20–60 characters compared to the agent's multi-sentence output. This aligns with findings in other automated counseling domains where users tend to be more concise than the agents guiding them.

The retrieval-based baselines (NAOMI-RAG and RAG+) exhibited high verbosity, with agent turns frequently exceeding 500 characters and peaking at over 1,200 characters. In contrast, NAOMI-DCA maintained a consistent and controlled response length, averaging approximately 160–250 characters per turn. This indicates that the DCA architecture successfully regulated the "lecture mode" often observed in RAG systems, maintaining a conciseness comparable to the prompt-based baseline (PT) while preserving the system's ability to steer responses to align with the goals of MI processes.

G Ablations

We conducted two sets of ablations for NAOMI-DCA. The first analyzes the local code-selection controller through offline counterfactual replay on collected DCA transcripts. The second studies the dynamic prompt used after code selection, isolating how MI code-definition specificity in the system prompt affects realization of an intended counselor code. All code-selection ablations were performed only on NAOMI-DCA transcripts, since this version tracks both explicit MI stages and controller-selected counselor codes. In total, the code-selection analysis covers 161 counselor turns.

ID	Question Text
Q1	I have heard of AI counselors before now.
Q2	I know people who have thought about using AI counselors.
Q3	I want to try AI counseling before using it for an extended time.
Q4	I have tried other conversational AI agents (e.g. chatbots or ChatGPT).
Q5	I have a positive attitude about conversational AI agents (e.g. chatbots or ChatGPT).
Q6	I know people who have tried other conversational AI agents (e.g. chatbots or ChatGPT).
Q7	Human counselors can help people lose weight.

Table 13: Questions in the pre-interaction acceptability questionnaire

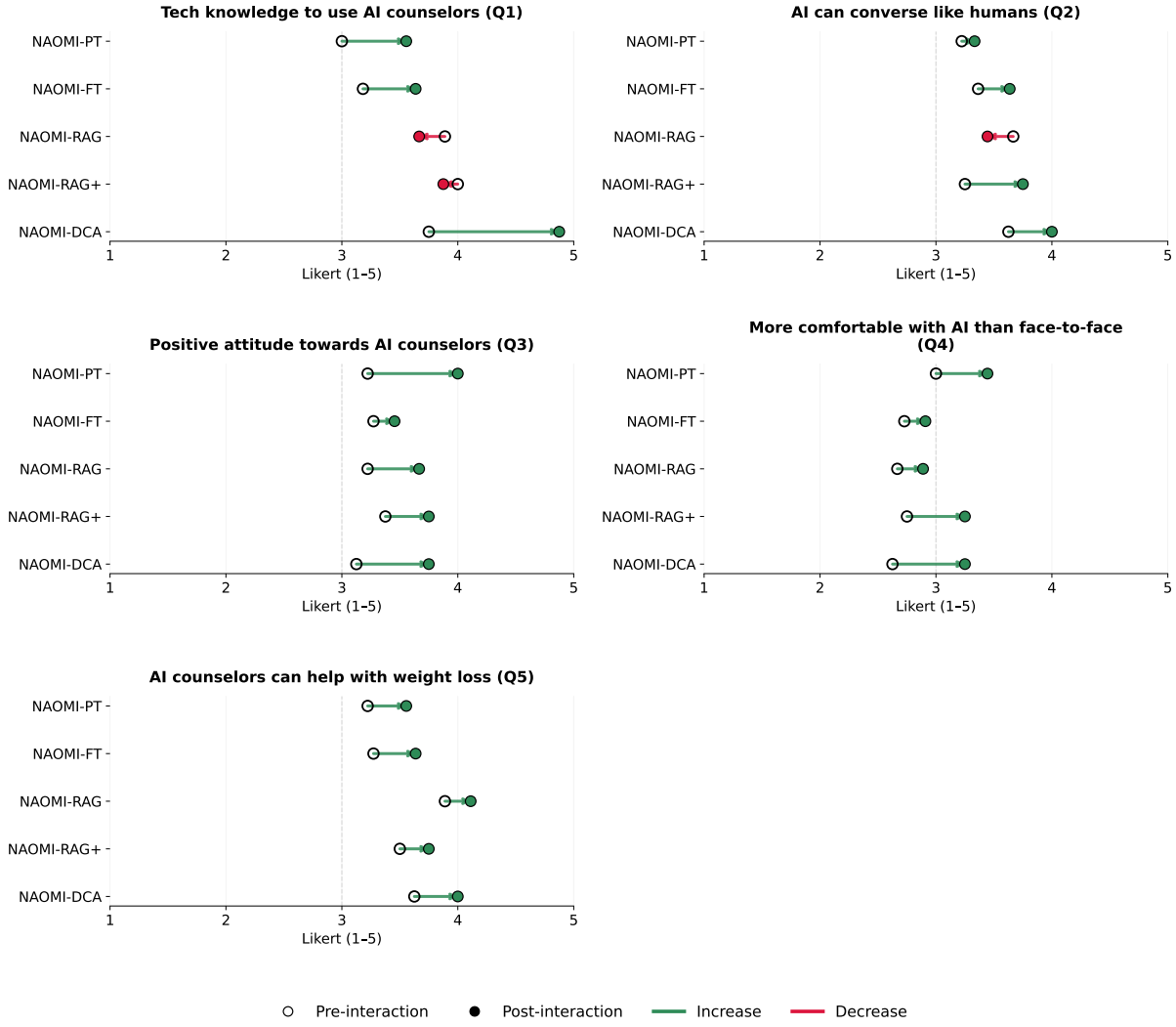

Figure 5: Changes in participant attitudes on the five repeated acceptability measures (Likert 1–5) from the questions administered pre- and post-interaction (Table 12) across different versions of NAOMI. Hollow markers indicate baseline (pre-interaction) means; solid markers indicate post-interaction means. Arrows highlight the direction and magnitude of the shift for each version of NAOMI.

G.1 Offline Replay Setup and Metrics

To analyze the local controller without rerunning full dialogue generation, we implemented a deterministic offline replay version of the code-selection module. At each counselor turn, the offline scorer reconstructs the candidate set from the current MI stage and the counselor code history, then recomputes candidate scores using the same components as the deployed controller: (i) pattern adherence, (ii) alignment with the expert-specified stage target distribution, and (iii) the Gated Recurrent Unit (GRU) prior for the next counselor code. The MI

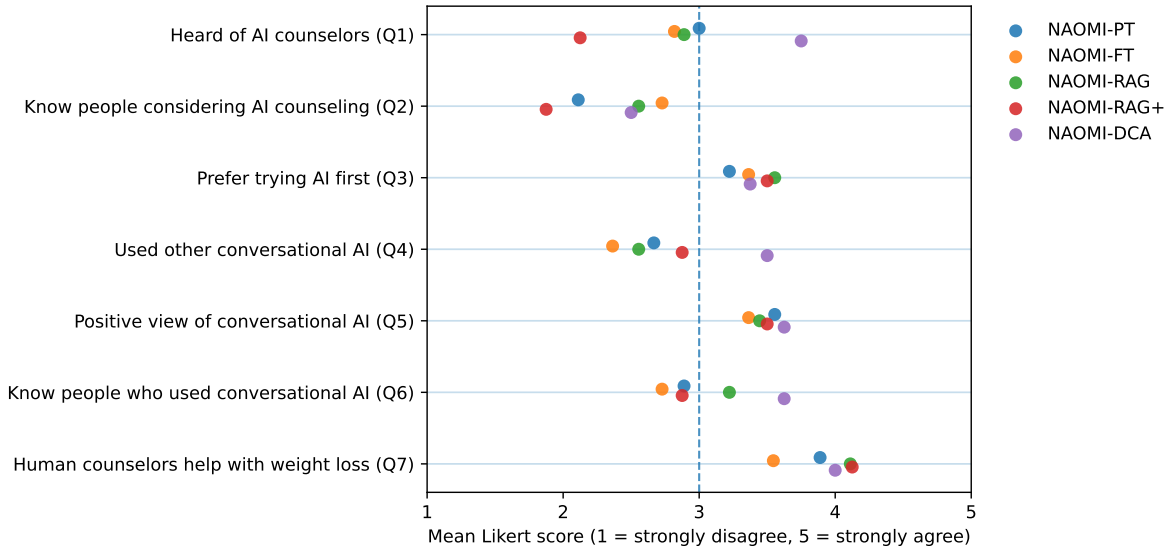

Figure 6: Distribution of participant responses (Likert 1–5) to the questions in the Pre-Interaction Acceptability Questionnaire (Table 13) across different versions of NAOMI.

Construct	Item Text
Importance	On a scale from 1 to 10, how important is it to you right now to lose weight?
Confidence	On a scale from 1 to 10, how confident are you that you would succeed at losing weight if you started now?
Readiness	On a scale from 1 to 10, how ready are you to start making changes right now to lose weight?

Table 14: Motivation ruler items measuring behavior change precursors administered pre-, post- and one week after the interaction

inconsistent codes blacklist filter can be enabled or disabled for counterfactual analysis.

For a given ablation, let $\hat{c}_t^{\text{base}}$ denote the code selected by the baseline replay at turn t, and let $\hat{c}_t^{\text{abl}}$ denote the code selected under the ablated setting. We quantify turn-level instability using code change rate (CCR):

$$\text{CCR} = \frac{1}{N}\sum_{t=1}^{N} \mathbb{1}\left[\hat{c}_t^{\text{abl}} \neq \hat{c}_t^{\text{base}}\right]. \qquad (1)$$

To assess whether a replayed code sequence preserves the intended stage-level behavioral profile, we also compute stage-wise KL divergence between the empirical replay distribution and the expert-specified target distribution for each MI stage:

$$\text{KL}_s = \sum_{c \in \mathcal{A}_s} \hat{p}_s(c) \log \frac{\hat{p}_s(c)}{\tilde{\pi}_s(c)}, \qquad (2)$$

where $\mathcal{A}_s$ is the stage-valid code support, $\hat{p}_s$ is the smoothed empirical replay distribution for stage s, and $\tilde{\pi}_s$ is the expert target distribution renormalized to the same support. For replay validation, the reference labels are the codes logged by the deployed controller during the original sessions. For the weight-sweep analysis, the stage target distributions $\tilde{\pi}_s$ serve as the gold reference for distributional alignment.

G.2 Code-selection Ablations

Before making counterfactual claims, we first verified that the offline scorer reproduces the logged controller closely enough to support ablation analysis. Under the deployed controller setting with the blacklist enabled, the offline replay matches the logged controller on 145 of 161 counselor turns, corresponding to 90.1% replay accuracy. This agreement is sufficiently high to support replay-based analysis, while the remaining mismatches likely reflect residual stochasticity in the original controller decisions when pattern score is a binary number, and KL divergence is the same for a set of MI codes in the same code category (Appendix A.3.1), with final decision being up to the GRU component.

We next estimate the degree to which each scoring component contributes to turn-level code selection. Starting from the production replay setting, we remove one controller term at a time—pattern adherence, distribution alignment, or GRU prior—while keeping the blacklist enabled and renormalizing the remaining weights to preserve their relative contribution. We then compare the resulting replayed code sequence to the baseline replay.

ID	Question Text
Q1	I like that counseling with AI agents is anonymous.
Q2	My personal information would be safe when using AI counselors.
Q3	AI agents are an acceptable way to receive counseling.
Q4	AI agents would save time traveling to a specialist or a clinic for face-to-face counseling.
Q5	Counseling with AI agents would be cheaper than face-to-face counseling.
Q6	Using AI agents can improve my access to counseling services.
Q7	Using AI counselors fits well with my lifestyle.
Q8	Using AI counselors fits well with my beliefs about counseling.
Q9	Using AI agents fits well with the way I like to receive counseling.
Q10	I could see myself using AI counselors in the future.
Q11	I could see myself using AI counselors for an extended period of time.

Table 15: Questions in the post-interaction acceptability questionnaire

#	Question Text
Q1	I believe that NAOMI could help people like me lose weight.
Q2	NAOMI met my weight loss counseling needs.
Q3	The way I interacted with NAOMI was pleasant.
Q4	I liked using NAOMI.
Q5	NAOMI is simple and easy to understand.
Q6	NAOMI is able to do everything I would want it to be able to do.
Q7	I felt that I was able to express myself effectively when using NAOMI.
Q8	When communicating to NAOMI, I felt as if I was communicating to a human counselor.
Q9	I think that NAOMI is as effective as a human counselor.
Q10	NAOMI's messages were well-formed.
Q11	NAOMI's messages were understandable to me.
Q12	NAOMI's messages were appropriate for a counselor.
Q13	I felt comfortable communicating with NAOMI about weight loss.
Q14	I would use NAOMI for weight loss counseling again.
Q15	Overall, I am satisfied with NAOMI.

Table 16: Questions in the usability questionnaire

#	Question Text
Q1	(How was NAOMI at...) Making you feel at ease (being friendly and warm toward you; treating you with respect; not cold or abrupt).
Q2	Letting you tell your "story" (giving you time to fully describe your issues in your own words; not interrupting you).
Q3	Really listening (paying close attention to what you were saying).
Q4	Being interested in you as a whole person (asking relevant details about your life; not treating you as "just a number").
Q5	Fully understanding your concerns (communicating that it had accurately understood your concerns; not overlooking or dismissing anything).
Q6	Showing care and compassion (seeming genuinely concerned; connecting with you on a human level; not being indifferent or detached).
Q7	Being positive (having a positive approach and a positive attitude; being honest but not negative about your problems).
Q8	Explaining things clearly (fully answering your questions; explaining things clearly; giving you adequate information; not being vague).
Q9	Helping you to take control (exploring what you can do to improve your health; encouraging rather than lecturing).
Q10	Making a plan of action with you (discussing options; involving you in decisions as much as you want; not ignoring your views).

Table 17: Questions in the CARE empathy questionnaire

We also test whether the stage-specific MI code filtering (blacklisting) actively constrained code selection in practice. For each counselor turn, we rescore candidates twice under the same weights: once with the blacklist enabled and once with it disabled. This yields two related quantities: (i) *blacklist pressure*, defined as the proportion of turns for which the unconstrained top-ranked code is forbid-

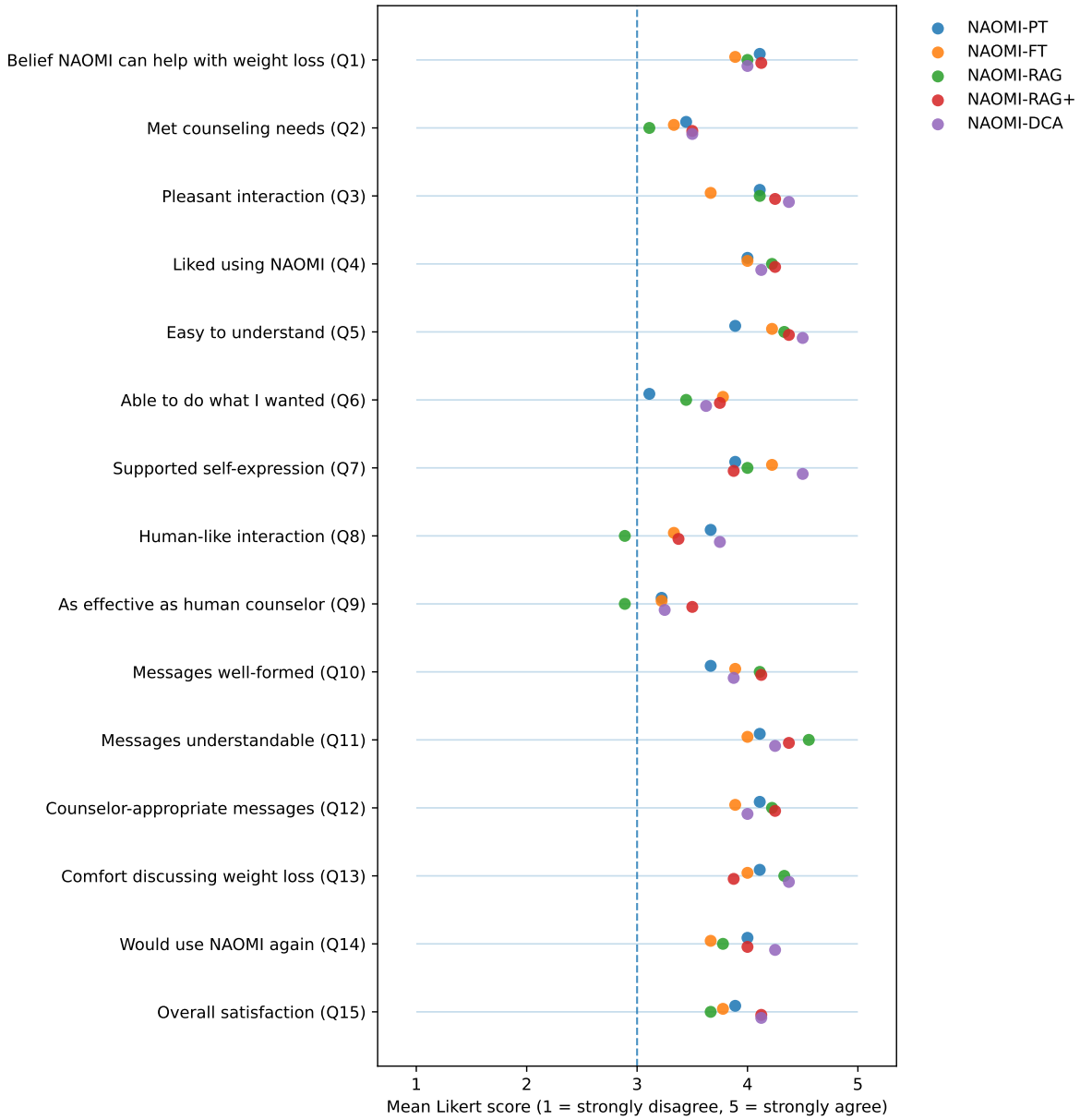

Figure 7: Distribution of participant responses (Likert 1–5) to the questions in the Usability Questionnaire (Table 16) across different versions of NAOMI.

den in the current stage, and (ii) *decision change rate*, defined as the proportion of turns for which enabling the blacklist changes the selected code.

Table 18 shows that removing any single controller term substantially changes the replayed decision sequence. The largest effect is observed when removing the distribution term (35.4%), followed by the pattern term (31.1%) and the GRU prior (29.2%). The three values are of similar magnitude, indicating that no single component dominates controller behavior. Instead, the local controller relies on these signals in a materially complementary way.

The blacklist results show that, on this corpus, the blacklist was non-binding: the unconstrained winner was never blacklisted for its stage, and enabling the blacklist never altered the final decision. Thus, under the observed logged sessions, blacklist constraints did not contribute additional filtering beyond what was already enforced by the other controller components. We believe that this is due to training data containing no blacklisted MI inconsistent codes.

G.3 Weight-sweep Sensitivity

To study sensitivity to controller weighting, we swept the simplex of relative weights assigned to pattern adherence, distribution alignment, and GRU

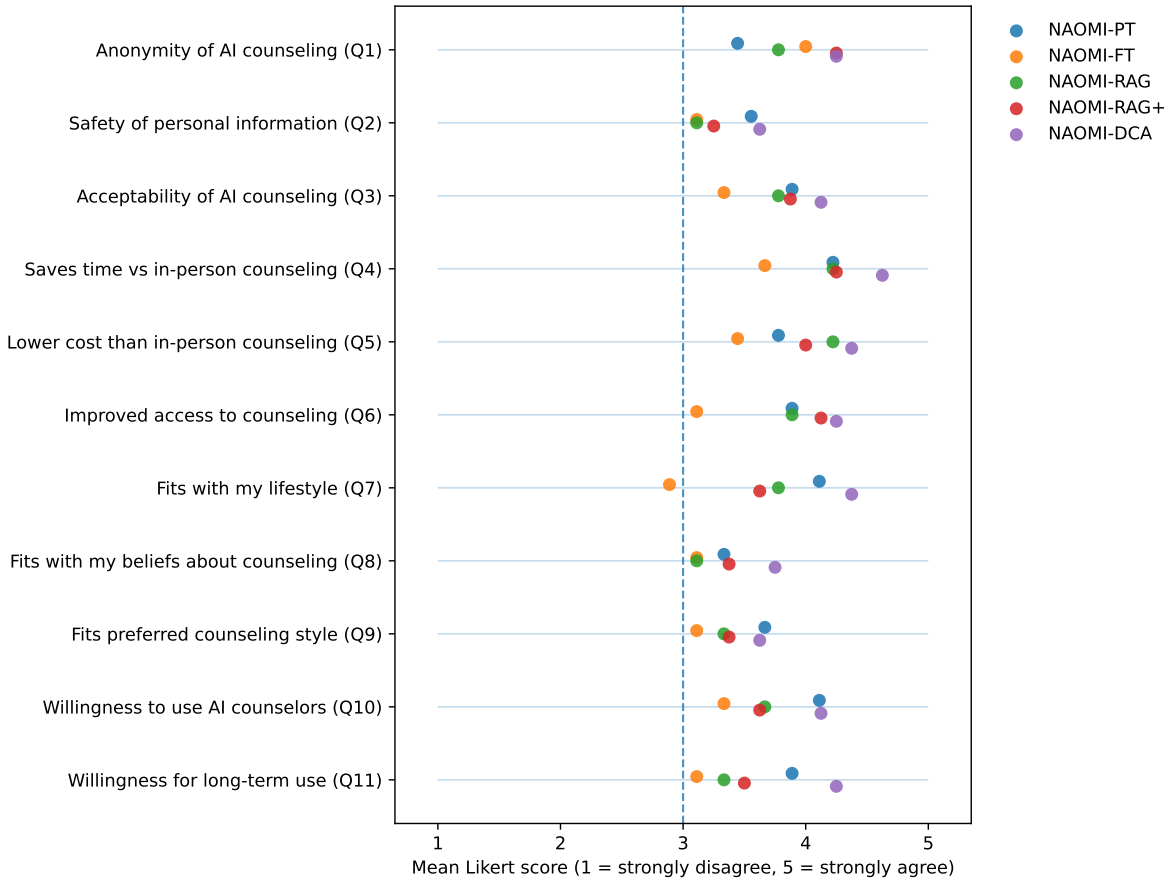

Figure 8: Distribution of participant responses (Likert 1–5) to the questions in the Post-Interaction Acceptability Questionnaire (Table 15) across different versions of NAOMI.

prior using a step size of 0.05, while keeping blacklist constraints enabled. For each weight setting, we replayed controller decisions over all logged turns and computed two aggregate metrics: (i) CCR relative to the baseline replay and (ii) stage-wise KL divergence from the expert-specified target stage distributions. The former captures turn-level decision instability, while the latter captures stage-level behavioral drift from the intended MI profile.

The deployed NAOMI-DCA controller used weights $(0.70, 0.25, 0.05)$ for pattern, distribution, and prior, respectively. These weights were chosen a priori to emphasize expert-authored communication patterns. However, when optimizing the offline distributional objective, the lowest-KL setting on the explored grid was $(0.40, 0.55, 0.15)$, which shifts substantially more weight toward the distribution alignment term.

This result is informative in two ways. First, it confirms that distribution alignment is not redundant: the offline optimum for preserving the expert stage-level code mix assigns it the largest weight. Second, it should not be over-interpreted as evidence that the deployed weights were mis-specified, since the production system intentionally prioritized expert-defined communication patterns over purely distributional matching. In other words, the sweep identifies a post hoc optimum for a specific offline objective, whereas the deployed setting reflects a design preference for more deterministic expert-controlled structure.

G.4 Dynamic Prompt Ablation

Beyond code selection itself, we also evaluated how an MI code definition affects the model's ability to realize an intended counselor communication behavior behind the code in the generated response. Using logged session contexts, we generated responses under three prompt variants while holding constant the MI stage, recent stage-specific context, latest client utterance, and target counselor code. The three variants were: *ND* (*no_defs*), which provides only the target counselor code label; *AD* (*all_defs*), which provides definitions for all counselor codes valid in the current MI stage; and *SO*

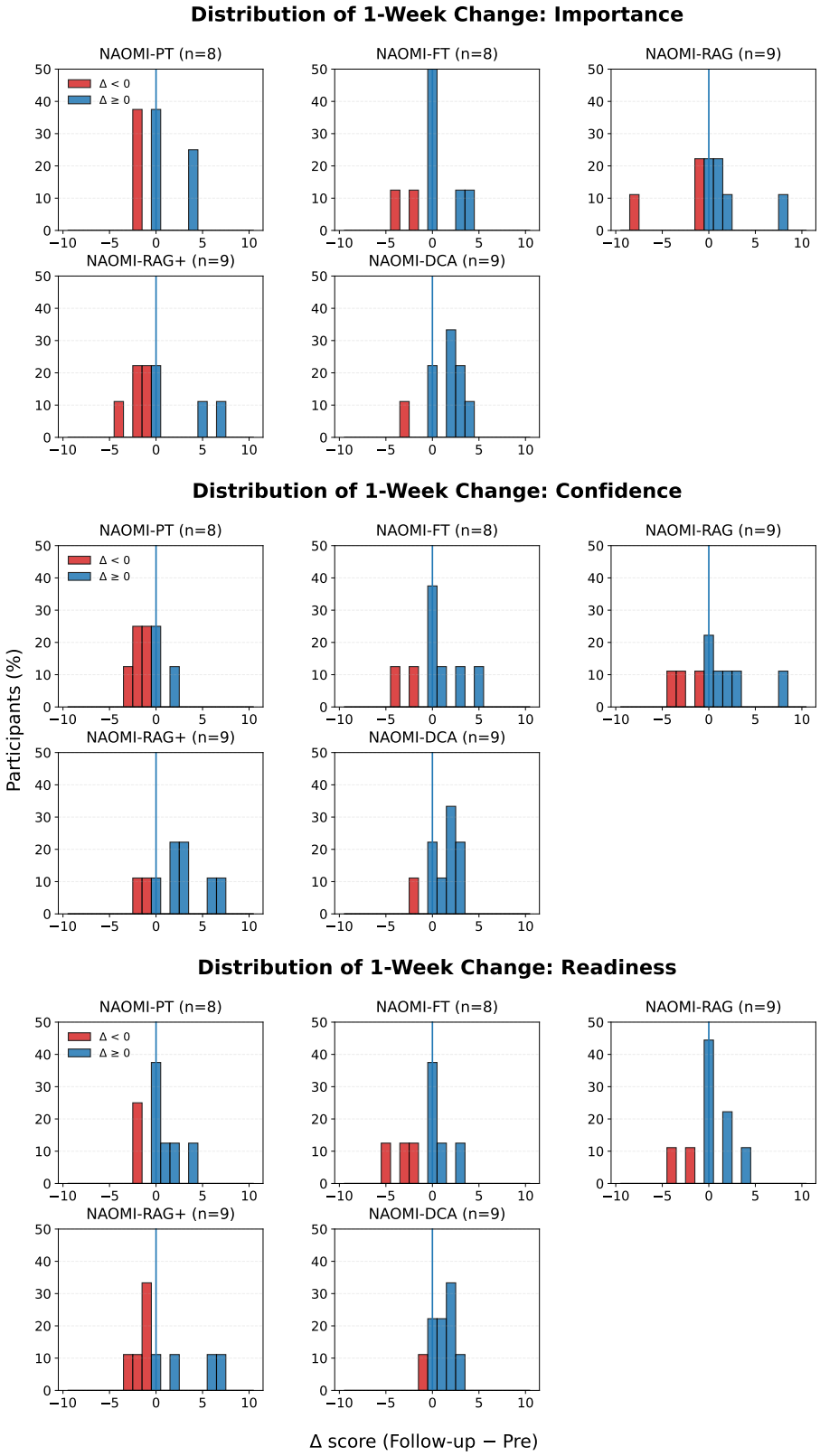

Figure 9: Distribution of 1-week changes (follow-up relative to pre-interaction) in the motivation ruler scores (Table 14) across different versions of NAOMI. Results are shown for Importance, Confidence, and Readiness, with red bars indicating the percentage of participants exhibiting negative and blue indicating positive or no change.

Analysis	Metric	Result
Replay validation	Replay accuracy vs. logged controller	90.1%
Remove distribution term	CCR vs. baseline replay	35.4%
Remove pattern term	CCR vs. baseline replay	31.1%
Remove GRU prior	CCR vs. baseline replay	29.2%
Blacklist ablation	Decision change rate	0.0%

Table 18: Summary of offline code-selection ablations on logged NAOMI-DCA counselor turns. Replay fidelity is high enough to support counterfactual analysis, and removing any one controller term causes substantial decision changes. By contrast, the stage blacklist is non-binding on the observed corpus.

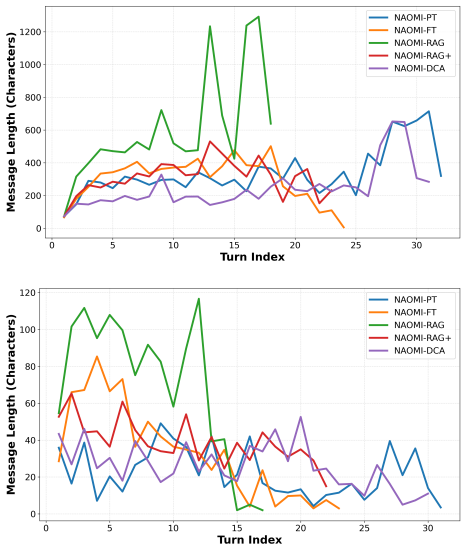

Figure 10: Mean response length (in characters) per turn across different versions of NAOMI. Top panel: agent response lengths. Bottom panel: client response lengths.

(*selected_only*), which provides only the definition of the intended target code. An example MI code definition is provided in Appendix B.6.

The generated responses were then scored by three independent LLM judges—Gemini 3.1 Pro (Google DeepMind, 2026), GPT-5.4 Thinking (OpenAI, 2026), and Llama 3.1 70B (Grattafiori et al., 2024)—using the same evaluation prompt and scoring rubric. For the proprietary judges, Gemini 3.1 Pro and GPT-5.4 Thinking, we used their default configurations through the Web interfaces and provided the full relevant prior conversational context needed for evaluation. All judges were instructed to evaluate only communicative realization of the intended counselor code, ignoring general fluency, empathy, or overall helpfulness unless these directly affected code realization. The prompt template used for all judges is shown below, with placeholders indicating fields filled from the logged session context and generated response.

LLM Judge Prompt for MI Code Realization

Role. You are an expert evaluator of Motivational Interviewing (MI) therapist behaviors.
Task. Score how well the therapist response realizes the **INTENDED** therapist MI code.
Evaluation focus. Judge only the *communicative function* of the response.
Scoring rubric.

1. clearly not the intended code
2. mostly another code / weak mismatch
3. mixed or ambiguous
4. mostly intended code but not perfect
5. clearly and strongly realizes the intended code

Rules.

- Do not judge general helpfulness or empathy.
- If another code fits better, assign 1 or 2.
- If the response is mixed, vague, generic, or unclear, assign 3.
- If the intended code is the best label but expressed imperfectly, assign 4.
- If the intended code is clearly and strongly realized, assign 5.
- If uncertain, choose the lower score.

Output format. Return **only** valid JSON in the following schema:

```
{
"match_score": 1-5,
"confidence": 1-5,
"reasoning_short": "<brief explanation>"
}
```

Inputs.

- `Stage: {STAGE}`
- `Intended code: {TARGET_CODE}`
- `Definition: {TARGET_CODE_DEF}`
- `Context: {CONTEXT}`
- `Response: {RESPONSE}`

The results are consistent across all three judges.

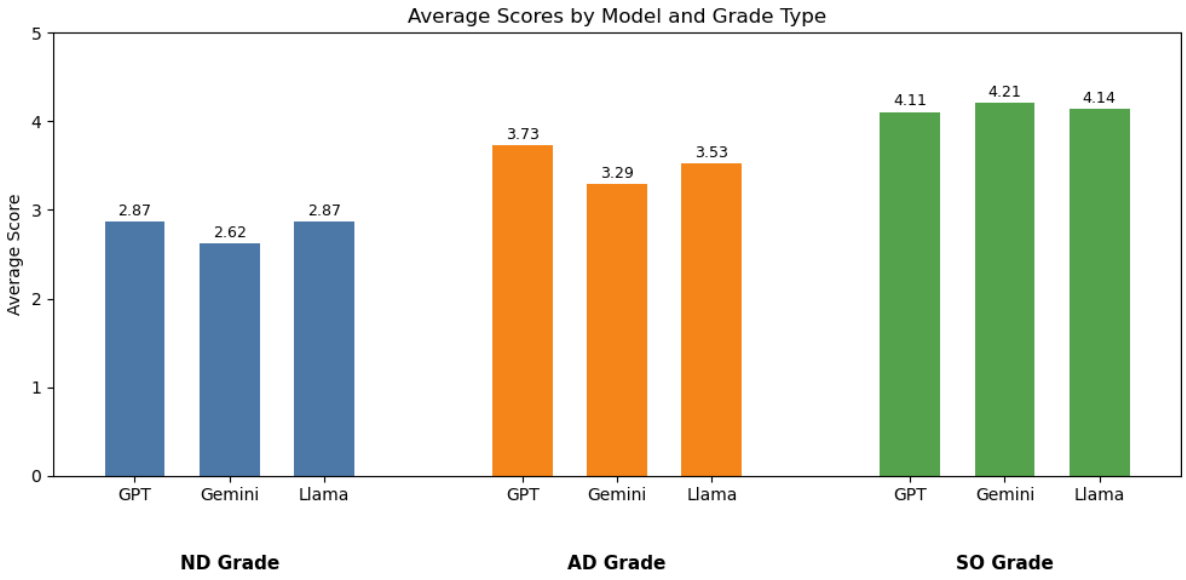

Figure 11: Mean code realization scores (1-5 range) under the three dynamic-prompt conditions: ND (*no_defs*), AD (*all_defs*), and SO (*selected_only*). Scores are averaged over the three LLM judges; higher is better.

Judge	ND	AD	SO
GPT-5.4 Thinking	2.869	3.735	4.107
Gemini 3.1 Pro	2.621	3.289	4.208
Llama 3.1 70B	2.869	3.527	4.144
Mean over judges	2.786	3.517	4.153

Table 19: Mean code realization scores (1-5 range) for the dynamic-prompt ablation. Higher values indicate better realization of the intended counselor code per score given by the LLM judges

As shown in Table 19, SO achieves the highest mean score overall (4.15), followed by AD (3.51), with ND substantially lower (2.78). Moreover, SO is the best-performing condition for all three judges individually. Relative to ND, adding definitions for all stage-valid codes improves the mean score by 0.730 points, while providing only the intended code's definition improves it by 1.36 points. Directly comparing the two definition-based settings, SO outperforms AD by 0.63 points on average.

These findings suggest that code definitions improve their realization, however a *targeted* definition is more effective than supplying a larger set of candidate definitions. One possible interpretation is that the prompt only with the selected code definition sharpens the intended communication target, whereas presenting all stage-valid definitions dilutes the signal by introducing competing code descriptions into the generation context.

H Implementation Details

H.1 Corpus of MI Session Transcripts with Human Counselors

To support supervised fine-tuning and retrieval-augmented generation, we curated a corpus of English-language motivational interviewing session transcripts focused on obesity counseling[2]. The dataset aggregates transcripts from two prior studies, comprising 154 counseling sessions in total. The corpus contains 50,720 dialogue turns (∼860,000 tokens), with an average session length of 329 turns. Expert counselor utterances account for 32,649 turns (64.4%) of the total. For the RAG-based architectures, this corpus served as the retrieval source for expert counselor demonstrations. For NAOMI-FT, it served as supervised fine-tuning data with MI behavior codes omitted, whereas for NAOMI-DCA, MI behavior codes were retained.

The two studies are a pilot study and a subsequent randomized clinical trial of an MI-based weight-loss intervention, both consisting of dyadic counseling sessions between a human MI counselor and a client that were audio recorded and subsequently transcribed.

The transcripts were manually annotated using MY-SCOPE (Carcone et al., 2013), an adaptation of the MI-SCOPE coding system (Martin et al., 2005) developed by an interdisciplinary team of

[2]will be shared with any interested party upon execution of data use agreement

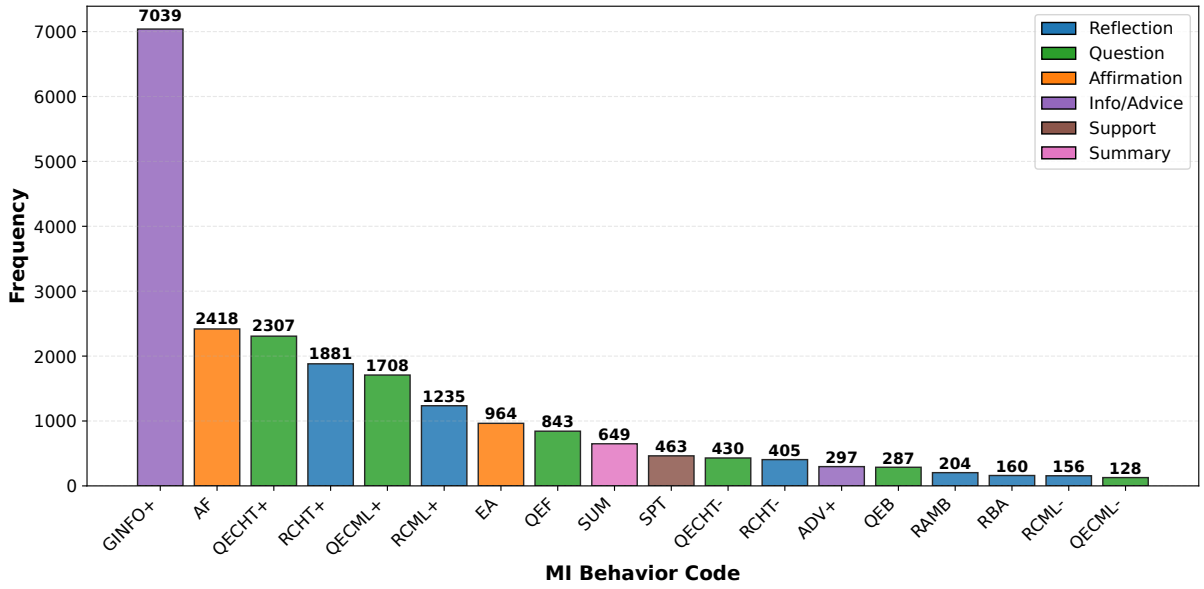

Figure 12: Distribution of expert counselor communication behaviors in the training corpus. The dataset reflects high-fidelity MI practice, dominated by Reflections (blue) and Questions (green).

MI trainers, a communication scientist, a linguist, and a community health worker, drawing on MISC (Miller et al., 2003), MITI (Moyers et al., 2016), and the broader MI literature. A portion of the sessions was co-coded to monitor agreement, which was substantial, with Cohen's κ ranging from 0.70 to 0.78.

To prepare the annotated transcripts for fine-tuning and retrieval, we converted the source documents to plain text, corrected artifacts introduced by transcription, and segmented each transcript so that every utterance occupies a single line together with the speaker label (Counselor or Client) and the corresponding communication behavior code.

We then consolidated the counselor and client codes. Simple and complex variants of each reflection type were combined, so that RCHT+S and RCHT+C both became RCHT+, and open and closed variants of each question type were combined in the same manner, with open-ended (OQECHT+) and close-ended questions to elicit change talk (CQECHT+) becoming QECHT+. The graded intensity levels used for client language were collapsed into positive and negative categories, for example, CHT+1 through CHT+3 became CHT+ and CHT-1 through CHT-3 became CHT-. This reduces roughly one hundred distinct annotation codes across the two data sources to the 18 counselor codes in Table 7 and the 7 client codes in Table 8. Summarization (SUM) was excluded from the set available to the local controller, since it is triggered by the global controller only at stage transitions.

Two groups of utterances were then removed. The first consists of MI-inadherent behaviors such as confronting, warning, and unsolicited advice (RC, CON, ADV-, GINFO-), which account for about 3% of the coded utterances. The second and considerably larger group consists of utterances coded with rare codes, such as session structuring, scheduling, or discussion unrelated to the target behavior (S-O, SS, SCH) and the TBN question codes. Figure 12 shows the distribution of counselor behavior codes in the resulting corpus.

H.2 Hardware Setup

Backbone LLM was fine-tuned and all versions of NAOMI deployed on a high-performance computing node equipped with 2× NVIDIA H100 PCIe GPUs (80GB VRAM each), a 16-core CPU, and 512 GB of RAM.

H.3 Fine-tuning Hyperparameters

We fine-tuned the Llama-3.1-70B-Instruct model (4-bit quantized) using the QLoRA fine-tuning method. We used the LoRA rank of $r = 8$ and alpha $\alpha = 16$ with a dropout of 0.05. Training was conducted for 300 steps using the AdamW 8-bit optimizer with a learning rate of 5×10^{-6} and a cosine scheduler (75 warmup steps). The effective batch size was set to 8. Complete hyperparameter details are listed in Table 20. For inference, we used nucleus sampling with $p = 0.95$ and $k = 40$, a temperature of 0.7, and a repetition penalty of 1.2

to ensure diverse and non-repetitive generation.

H.4 GRU-based Counselor Code Prior

For the learned prior, we use a lightweight Gated Recurrent Unit (GRU) model trained on code-only sequences derived from the code-annotated MI corpus described in Appendix H.1. We adopted a GRU as a simple sequential model for capturing short-range dependencies between counselor behavior codes while adding negligible parameter overhead compared to the backbone transformer (approximately 116K parameters in our final configuration).

Training instances are constructed with a sliding window over counselor code sequences: the input is the previous 10 counselor codes and the target is the next counselor code, with shorter histories left-padded using a special `<START>` token. In total, this produced 15,462 training instances over a vocabulary of 26 symbols.

The final model uses a 2-layer bidirectional GRU with hidden size 64, embedding size 32, and dropout 0.3, followed by a linear prediction layer over the counselor code vocabulary. Training uses Adam with class-weighted cross-entropy, along with a validation split and early stopping.

At inference time, the GRU operates only on recent counselor code history, not on raw text. Given the current session history, it produces a probability distribution over candidate counselor codes, which is used as the learned prior component of the local controller.

H.5 Web User Interface

NAOMI's Web interface, shown in Figure 13, features a microphone button for voice input and also allows a user to type responses into the text box.

I Sample Conversation Between NAOMI-DCA and a De-identified Client

This section presents a representative de-anonymized transcript illustrating NAOMI-DCA's progression through the four MI phases in one of the counseling sessions.

Counselor: Nice to meet you, `<participant name>`. What do you think about your weight?

Client: i just need to tone up, im finally losing weight

Counselor: You're finally seeing the scale move, which is a huge victory! Your focus is shifting from just shedding pounds to actually shaping and strengthening your body.

Client: yes justmaking sure my muscles arent flabby

Counselor: You want to feel solid and firm rather than just smaller. Building that muscle definition is what's really going to make you feel comfortable.

Client: yes health is everything

Counselor: Health is your absolute baseline; everything else in life builds right on top of that. What does being truly healthy mean for you personally?

Client: being healthy means , living longer and feeling great , and being able to function

Counselor: It's about longevity and having the vitality to actually enjoy those years. You want a body that works dependably for you, giving you the freedom to function without limits.

Client: yes i feel great about the positive changes in my life

Counselor: You're riding a wave of success right now, and that's feeding directly into your confidence. What do you think has been the secret ingredient driving these positive changes?

Client: positive vibes and great motivation

Counselor: You've tapped into a really powerful internal energy source. That upbeat, motivated mindset is the engine driving all of your progress.

Client: yes it feels good

Counselor: It feels incredibly rewarding to be in the driver's seat of your own health. What specific actions have you been taking to harness those positive vibes and lose the weight?

Client: just eating less junk and walking more

Counselor: You've simplified it beautifully by improving your fuel and increasing your movement. Since you're making such great strides, what usually pops up as a barrier to keeping these habits going?

Client: mostly just getting tired after work

Parameter	Value
Base Model	Llama-3.1-70B-Instruct
Quantization	4-bit (NF4)
LoRA Rank (r) / Alpha (α)	8 / 16
Learning Rate	5×10^{-6}
Batch Size (Effective)	8
Max Steps	300
Optimizer	AdamW (8-bit)

Table 20: Hyperparameters used for fine-tuning.

Counselor: That post-work exhaustion is a very real wall that hits you hard. When that energy crash happens, how does your evening usually play out?

Client: i just end up watching tv and doing nothing

Counselor: Your body and brain are totally drained from the day, so sinking into the couch to watch TV is the most natural way to recharge.

Client: yea

Counselor: It makes complete sense to need that downtime after a long day. To summarize so far, you've built amazing momentum by eating better and walking, but you want to shift to firming up, even with that post-work fatigue. Which behavior would you like to start with? We could start with diet, physical activity, or another part of your lifestyle that you'd like to change.

Client: physical activity. the walking is good but it doesnt tone me

Counselor: Physical activity is the logical next step. Walking helped you lose the weight, but you know you need a different challenge to build that muscle. What type of activity are you leaning toward?

Client: prolly lifting weights at home

Counselor: Lifting weights at home brings the workout right into your living room. It's a highly efficient way to build that firmness without adding extra complications to your day.

Client: exactly i hate the gym anyway

Counselor: The gym just isn't your scene, so creating your own private setup is the perfect workaround. So, we've narrowed our focus down to physical activity, specifically lifting weights at home. Why do you want to start with lifting weights?

Client: cuz i want my musmcles to be firm and not flabby like i said earlier

Counselor: Having firm, strong muscles is your ultimate definition of success right now. On a scale from 0 to 10, with 0 being not important at all and 10 being a top priority, how important is it to you to start this weight routine?

Client: maybe an 8

Counselor: An 8 is incredibly high! You are heavily invested in making this happen. What makes it an 8 for you, rather than a lower number like a 4?

Client: cuz im already seeing results from the diet and i want to keep going. summer is comin

Counselor: Summer acts as a fantastic deadline pulling you forward. You've already proven you can get results with your diet, and you want to ride that winning streak right into the summer. On that same 0 to 10 scale, how confident are you that you can consistently lift weights at home?

Client: maybe a 6

Counselor: A 6 shows you definitely believe you can do it, but there's a bit of realistic hesitation mixed in there. What usually acts as a roadblock and pulls your confidence down to a 6?

Client: i get bored or it feels like too much work

Counselor: If a routine feels like a massive chore, your motivation naturally tanks. What would need to happen to

bump your confidence up from a 6 to an 8 and keep boredom out of the picture?

Client: if i had a simple plan. i dont want to work out for an hour

Counselor: Keeping the routine short and sweet is the secret to consistency for you. You want maximum results with a minimal time commitment, avoiding that hour-long slog.

Client: yeah exactly

Counselor: You've got the vision perfectly dialed in. To quickly summarize, you're driven to lift weights to firm up for summer, and keeping the workouts short is your key to staying confident. What are some steps you would consider taking to help you achieve your goal? What actions seem within reach this time?

Client: maybe doing a 20 min workout video on youtube a few days a week

Counselor: A 20-minute YouTube video completely solves the boredom problem and keeps it from feeling like too much work. How do you see us working together to fit those few days smoothly into your weekly schedule?

Client: tuesdays and thursdays before work. if i wait until after work ill just sit down

Counselor: Tuesday and Thursday mornings are your golden windows. You are outsmarting that post-work couch trap by getting the workout done before the day even begins. How will you make sure everything is ready to go when your alarm goes off?

Client: ill put my clothes out the night before and get the video out ready

Counselor: Laying out your clothes and queuing up the video the night before removes all the friction from your morning. Life does happen, though—what is your backup plan if you wake up on a Tuesday and you're just too exhausted to do the full 20 minutes?

Client: i guess i could just skip it and try wednesday

Counselor: Moving it to Wednesday is a solid safety net. Some people also find it helpful to do a "mini-version," like just 10 minutes of the video, so they keep the habit alive. It's totally up to you, but what do you think of a shorter backup like that?

Client: yeah 10 minutes is better than zeiro. I need to stop being all or nothing

Counselor: Dropping that all-or-nothing mindset is a massive breakthrough. You recognize that giving 10 minutes on a tough day is infinitely better than giving zero. How will you track or celebrate these wins once you finish a session?

Client: maybe just check it off on my calendar so i can see I did it

Counselor: Checking it off your calendar gives you instant, visual proof of your discipline. Let me summarize your plan: you're going to do a 20-minute YouTube strength video on Tuesday and Thursday mornings, prepping everything the night before. If you're tired, your backup is to just do 10 minutes to ditch the all-or-nothing mindset, and you'll track it on your calendar. On a scale of 0 to 10, how confident do you feel about this exact plan?

Client: a 9 now

Counselor: Jumping up to a 9 is phenomenal! You've built a highly realistic plan that bypasses the gym, fits your life, and uses the positive momentum you've already created to get firm for summer. To wrap up our session today: you're shifting from walking to a simple, 20-minute morning weight routine to tone up, utilizing a smart backup plan to keep you consistent.

Client: thanks me too

Counselor: You are very welcome! You have a solid routine, morning prep, and a brilliant 10-minute backup plan locked in. I'm really excited for you to step into summer feeling incredibly strong. Have a great week!

J Illustrative Failure Cases

Although the developed agents generally produced responses appropriate for an MI weight loss coun-

selor, this section presents representative examples when the agents did not produce the intended or expected response. These examples are included to illustrate important limitations that may not be fully reflected in the aggregate results of quantitative evaluation.

J.1 Failure Cases in Retrieval of Exemplar Counselor Responses for NAOMI-RAG

Retrieval-augmented generation was intended to ground model outputs in relevant exemplar counselor utterances from the MI transcript corpus. However, retrieval was not always sufficiently accurate. In some cases, the retrieved examples were only weakly related to the client's most recent utterance, overly generic, or introduced irrelevant context. This could cause the generated response to drift from the client's actual concern, become repetitive, or take a direction that was not well aligned with the immediate conversational context. A possible contributing factor is the limited size of the transcript corpus, which may have restricted the availability of closely matching exemplars.

Example failure case for NAOMI-RAG retrieval

Latest client query: *sticking with it*
Conversation context:
Client: how difficult it is
Counselor: It sounds like you're feeling a bit overwhelmed by the idea of losing weight. Can you tell me more about what specifically feels difficult to you? Is it the thought of changing your daily habits, or something else entirely?
Client: sticking with it
Retrieved counselor responses:

1. *And you see that that's one way that you can really work on stopping the eating when you're bored. Now you mentioned that your grandma really has kind of been tough on you about your weight.*
2. *What do you think you might do with it?*

In this example, the retrieved counselor responses are not well aligned with the client's immediate concern about difficulty maintaining weight-loss efforts. The first retrieved response introduces unrelated context about boredom eating and the client's grandmother, while the second is too generic to provide useful grounding for the current turn. As a result, the retrieved exemplars may steer the generated response away from the client's actual concern rather than helping the model respond in a focused and contextually appropriate way.

J.2 Failure Cases in DCA Responses Coded as Persuasion by MITI Coders

Although DCA was designed to promote stage-appropriate MI behavior through explicit code selection and dynamic prompting, some generated responses were nevertheless coded by MITI annotators as *Persuade*. These cases are important because persuasion is generally inconsistent with high-fidelity MI when it directs the client toward change without sufficient emphasis on autonomy or collaboration.

Example failure case for NAOMI-DCA response coded as persuasion

Client:: I might share my snack with them
Counselor: So you're thinking of sharing your healthy snacks with them—that's a great way to stay on track and also maybe even influence those around you to make some healthier choices too. How do you think they'll react to you sharing your snacks, and what might that do for your relationships with them?
Client:: It may motivate them to start making healthier choices
Counselor: You came into our conversation today because of concerns about your health and wanting to make changes to live a long and prosperous life. This desire is deeply connected to being there for your grandkids, which brings you joy and love, and motivates you to be healthy and energetic enough to watch them grow and help them develop into mature adults. Together, we explored how making healthier dietary choices could be a great starting point. You've committed to giving up sugary drinks and sweets, replacing them with water or juice and healthy snacks like cashews and grapes. Your plan includes keeping these options on hand at home, work, and even in your car or bag so they're always available when needed. When faced with social situations where others might be having unhealthy treats, you've thought about sharing your healthy snacks as a way to stay on track and potentially inspire those around you. Remember, taking care of your health is not just about the food you eat; it's also about showing up fully for yourself and your loved ones. I have no doubt that with this

drive and commitment, especially inspired by your grandkids, you will find success in your journey towards healthier living. Please let the research assistant know that our conversation has come to an end.

In this example, the highlighted sentence was coded by MITI annotators as *Persuade*. Although the response is coherent and remains broadly aligned with the client's goals, the phrasing moves beyond reflection and summary into more directive motivational language. Importantly, this behavior does not stem directly from the DCA response policy itself, but rather from the auxiliary stage summarization component (in this example, for the planning stage), whose generated summary have introduced wording that was then incorporated into the final response.

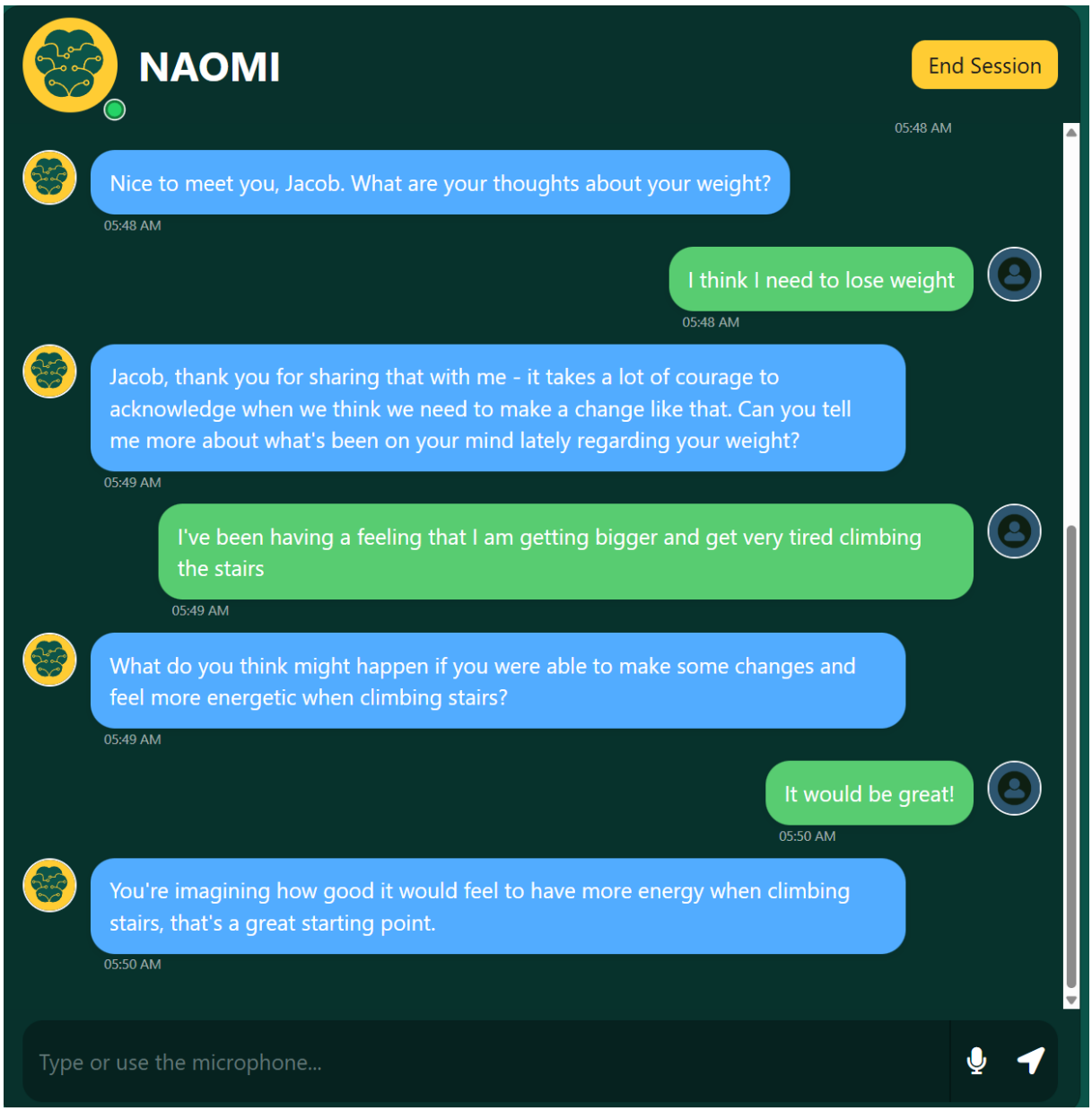

Figure 13: The NAOMI Web user interface. The screenshot demonstrates a session in progress, featuring the chat history view, session management controls in the header, and the text/audio input bar at the bottom.