Uncertainty-Aware Health Coaching for Sustainable Habit Building with Human Oversight

Motivation and Background

Health coaching reflects a shift from episodic, clinician-centered care toward continuous, personalized support that helps individuals build and sustain healthy habits in their daily lives. As a non-clinical practice, health coaching emphasizes goal setting, self-reflection, motivation, and lifestyle behavior change rather than diagnosis or treatment. An example excerpt from a health coaching conversation is provided in Annex A. A substantial body of evidence demonstrates the effectiveness of health coaching across diverse outcomes, including health behavior change, lifestyle modification, and goal attainment [1]. However, traditional health coaching depends on sustained human involvement, making it labor-intensive and difficult to scale while preserving consistency and long-term support. This limitation motivates the exploration of new delivery models that retain the core strengths of human coaching while extending its reach and continuity.

AI-based systems, particularly those built on large language models (LLMs), offer the potential to provide continuous, personalized coaching support at scale, reinforcing healthy habits and maintaining engagement between human interactions [2, 3]. While recent multi-session LLM-based coaching systems demonstrate basic personalization and longitudinal interaction, they remain limited by brittle memory, repetitive or rigid dialogue patterns, and—critically—the absence of uncertainty awareness, safety-aware reflection, and structured human oversight [4]. Existing approaches largely operate as static deployments, lacking mechanisms to signal uncertainty, escalate ambiguous situations to human coaches, or adapt through ongoing expert feedback [5]. As a result, current systems can generate supportive language but fall short of the trustworthy, adaptive human-AI collaboration required to balance scalability with safety in non-clinical health coaching.

Project Description

We propose to develop an uncertainty-aware health coaching system that addresses key limitations of existing approaches through three primary innovations: (1) specialized reflection agents that support safety monitoring, session planning, and learning from interaction histories; (2) a human-in-the-loop framework with progressive system autonomy, in which automated capabilities expand as reliability improves while human judgment remains available for safety-critical cases; and (3) reinforcement learning from human feedback that enables the system to improve through real-world deployment. The proposed system builds on established memory and conversation agents, focusing innovation on adaptive reflection mechanisms, explicit uncertainty estimation, and structured human-AI collaboration protocols.

Human involvement in the system serves two roles. First, humans provide real-time, turn-level monitoring and intervention to manage safety risks and high-uncertainty situations during coaching interactions. Second, humans conduct session-level reflective reviews to assess safety, goal alignment, and coaching quality, and to provide feedback for system improvement. The system is initially deployed with maximal human involvement and gradually reduces routine oversight as the model learns through uncertainty-aware decision making and reinforcement learning. This approach addresses key limitations of existing systems, including the lack of uncertainty modeling, adaptive autonomy, and structured incorporation of expert judgment, while ensuring that automated coaching remains within appropriate ethical and safety boundaries.

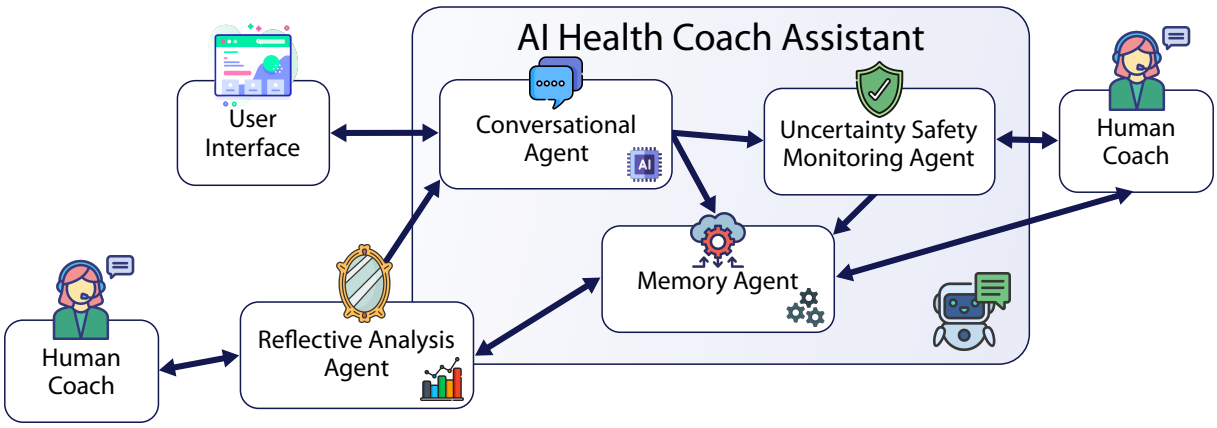

Figure 1: Proposed multi-agent architecture for uncertainty-aware health coaching with human-in-the-loop.

Figure 1 illustrates our proposed multi-agent system, which consists of a User Interface, Conversation Agent, Uncertainty and Safety Monitoring Agent, Memory Agent, and Reflective Analysis Agent. The Conversation and Memory Agents handle dialogue generation and client-specific data management, while the Uncertainty and Safety Monitoring Agent evaluates each response for harmful content, guideline violations, or off-topic drift. When safety concerns or high uncertainty are detected, the system immediately escalates to a human coach for real-time intervention; otherwise, sessions are reviewed offline after completion. The Reflective Analysis Agent aggregates session data, proposes guideline or prompt adjustments, and incorporates coach feedback into the system's learning loop, enabling adaptive prompt selection and progressive system competence. This architecture ensures continuous improvement, maintains safety, and balances automated coaching with human oversight.

Existing Dataset. CoachLah is a unique Singlish–English parallel corpus of real-world health coaching sessions collected during the first year of an ongoing randomized controlled trial. The current release comprises 86 sessions (12,114 utterances, 80+ hours) from 32 clients and three professional coaches in Singapore. Dialogues are speaker-labeled, feature natural code-switching in Singlish, and are aligned with English translations. The dataset captures authentic, culturally diverse coaching interactions and includes longitudinal annotations of 138 SMART behavior goals across 68 sessions, enabling analysis of goal evolution and training of uncertainty-aware, human-in-the-loop reflection agents. Participants represent Singapore's multicultural population, and sessions focus on education, goal setting, and accountability. Data collection is ongoing, and a substantially larger corpus is expected by Summer 2026.

We will leverage the CoachLah dataset by clustering real coaching conversations to identify representative interaction patterns and derive diverse scenarios for system testing and simulated interactions. These scenarios will support active learning and controlled data generation through interactions with the proposed multi-agent system. In parallel, existing human-human coaching sessions will be evaluated using structured checklists for guideline adherence, response quality, and safety, producing labeled data to train an initial uncertainty prediction model. Together, these steps enable safe, data-driven system development, strengthen reflection-agent performance, and provide a foundation for reinforcement learning from human feedback.

Project Outcomes and Deliverables

Our project advances both the scientific understanding of uncertainty-aware human-AI collaboration and the development of safe, scalable systems for health coaching.

1. **Open-source health coaching system:** Fully documented implementation of the multi-agent architecture with reflection agents, uncertainty estimation modules, progressive oversight protocols, and deployment guides, to be released on GitHub.
2. **Uncertainty-aware evaluation and prediction framework:** Reproducible methodology for uncertainty estimation and calibration, including uncertainty prediction metrics, calibration protocols, safety reflection prompts, and LLM-as-judge evaluation templates designed for integration into the coaching workflow.
3. **Adaptive human-AI collaboration and progressive oversight guidelines:** Protocols for coach-AI interaction across uncertainty levels, including threshold setting, dynamic escalation and de-escalation, reflection reviews, safety workflows, and phase-based oversight adaptation.
4. **Reinforcement learning (RL) training pipeline:** Open-source implementation of RL from human feedback during reflection, including reward modeling from coach corrections, uncertainty-weighted policy updates, and training infrastructure for continuous system improvement.

Organization

Team. The project team brings together expertise from academia and industry in agentic AI, uncertainty-aware learning, human–AI interaction, and digital health. Core members and domain experts are in Table 1. Students will be added upon application.

Core Members	**Domain Experts**
Iva Bojic, NTU Singapore	Kimia Ghobadi, Johns Hopkins University, US
Michael Tänzer, NTU Singapore	Kristina Gligoric, Johns Hopkins University, US
Mahnoosh Mehrabani, Interactions LLC, US	Qi Chwen Ong, University of Oxford, UK
Ali Dadgar, Interactions LLC, US	Damianos Karakos, RTX BBN Technologies, US
Srinivas Bangalore, Interactions LLC, US	Vasudha Varadarajan, CMU, US
Andy Khong, NTU Singapore	2+ health coaches from Singapore/US

Table 1: Project team structure.

Work-plan. A preliminary timeline is presented in Table 2, with monthly pre-workshop tasks and weekly activities during the workshop.

Pre-Workshop (Monthly Milestones)	**During Workshop (Weekly Milestones)**
M1. Curate and prepare de-identified conversational datasets. Define safety, uncertainty, and reflection benchmarks.	**W1–W2.** Calibrate agent behavior under human-gated control, validate uncertainty and escalation thresholds.
M2. Expert annotation of pilot data. Implement core prototype components, including monitoring and offline reflection pipelines.	**W3.** Validate human-in-the-loop workflows through live escalation to human coaches under low-confidence conditions.
M3. Run pilot experiments to validate reflection workflows across turn-, session-, and system-level signals.	**W4–W5.** Evaluate progressive autonomy under risk-conditioned control and assess post-hoc safety and workload reduction.
M4. Finalize prototype and establish human-only baseline metrics.	**W6.** Final evaluations. Synthesize results and prepare conclusions for final presentation.

Table 2: Preliminary timeline of pre-workshop and workshop activities.

References

[1] J. M. Olsen and B. J. Nesbitt, "Health coaching to improve healthy lifestyle behaviors: An integrative review," *American journal of health promotion: AJHP*, vol. 25, no. 1, pp. e1–e12, 2010.

[2] M. Jörke, S. Sapkota, L. Warkenthien, N. Vainio, P. Schmiedmayer, E. Brunskill, and J. A. Landay, "Gptcoach: Towards llm-based physical activity coaching," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–46, 2025.

[3] M. V. Heinz, D. M. Mackin, B. M. Trudeau, S. Bhattacharya, Y. Wang, H. A. Banta, A. D. Jewett, A. J. Salzhauer, T. Z. Griffin, and N. C. Jacobson, "Randomized trial of a generative ai chatbot for mental health treatment," *Nejm Ai*, vol. 2, no. 4, p. AIoa2400802, 2025.

[4] M. Jörke, D. Genç, V. Teutschbein, S. Sapkota, S. Chung, P. Schmiedmayer, M. I. Campero, A. C. King, E. Brunskill, and J. A. Landay, "Bloom: Designing for llm-augmented behavior change interactions," *arXiv preprint arXiv:2510.05449*, 2025.

[5] M. Ozolcer and S. W. Bae, "Offline policy evaluation of multi-turn llm health coaching with real users," *arXiv preprint arXiv:2510.17173*, 2025.

A Coaching Conversation Example

This dialogue excerpt represents approximately five minutes from a one-hour coaching session with a male client in his mid-50s participating in a cholesterol management program. The session is the third in a planned series of five to six monthly coaching interactions focused on sustainable behavior change to support medication adherence and cardiovascular health. In previous sessions, the client established goals related to regular physical activity and meal timing, with mixed adherence; session two revealed consistent morning walks (two to three times weekly) but difficulties maintaining evening meal schedules due to work responsibilities. The current session opens with a review of these existing goals before the coach guides refinement of the exercise target to support the client's emerging weight loss objective.

COACH: Then your exercise, you say you wanted to do more. The previous time I noted that you want to take your stairs instead of the lift, and then you want to walk in the morning three times a day for an hour. So after, after the discussion today, I also hear that there's many other things that you are doing lah. So for exercise, what is something that you'd like me to note down and check in the next month?

CLIENT: Hmm. Okay, currently I'm doing about two times per week on my own and one time with the family. Uh, that is on, on the day when my daughter has a swimming lesson. Personally, if I can, I would... if I can, if I can increase to three to four times a week with a goal, with a goal of losing weight as well. So it come, yeah. So if I want to lose weight, I have to do something about it. I have to watch... I have to increase my exercise, I have to cut down my food timing intake, I have to sleep early as well. So probably these are three things that ultimately help me to reach my goal lah.

COACH: Okay, so for the exercise, now I'm hearing the... to increase your, to increase the days that you are doing it. So is it walking only, or is there other things that you're doing? So now I'm hearing walk three to four times a day. Am I noting it right?

CLIENT: Yeah, I think it could be... it would be walking or maybe going to the gym as well.

COACH: Okay.

CLIENT: Yeah. So it's a mixture of both lah.

COACH: Okay, so walk or gym lah, three to four times a day. Then every time when you are walking or at the gym, how long would that take? Or what intensity would that be? Because now you are trying to lose the weight, right? So in terms of the timing, the intensity, all these also plays, plays a part.

CLIENT: So if I were to walk, I would normally walk into the, the park. The park has up climbing inclination. So I'll probably walk about three to four rounds. Three to four rounds around the park going up, down, up and down. That's one. If I were to on a gym, I'll be on the treadmill. I'll increase the inclination level as well.

COACH: Hmm, okay.

CLIENT: Yeah. So I mean, I mean one of my friends had shared with me is that if you go to treadmill and then walk on a flat, it doesn't really help. So you

have to incline, incline it so that you... it helps you to burn more, burn more and lose more weight faster lah. Yeah, yeah.

COACH: So you're increasing the intensity to make it from a maybe light intensity to a more moderate intensity lah, when you are doing your walks. Okay, and for doing your walks with a higher intensity for three to four times a day, what would your confidence level be for the next month?

CLIENT: Probably seven.

COACH: Okay. How come seven? Not six?

CLIENT: I think I can do it now with the current routine that I have. I can get, get my wife to join me and we can adjust certain routine. Yeah. I think we can do that.

COACH: Hmm, so it's again bringing in, bringing the family in, making it like a, a couple activity that you can do with your wife. So not only you get healthier, your family gets healthier together with you.

CLIENT: Yeah.

COACH: Okay, so that's for the, for the exercise. And then the last one is the... oh, exercise, anything else you want to add inside? Or anything I've missed out?

CLIENT: So I've gym, walking, swimming... I think that's pretty much for now.

COACH: Okay.

CLIENT: That's it, yeah.

COACH: Okay, so that's... really as compared to three months ago where the activity levels were lesser, now you have gym, you have walking, you have swimming. It's really like a progress across the, across time.

B Analysis of the Coaching Dialogue

This dialogue excerpt illustrates goal refinement for exercise behavior, demonstrating how the coach guides the client from an initial two sessions per week to a target of three to four sessions weekly with increased intensity. The interaction exemplifies several core coaching techniques aligned with the structured goal-setting agenda.

The coach employs a **confidence ruler** assessment, eliciting a rating of seven out of ten. Notably, the coach probes this relatively high confidence level with "How come seven? Not six?" to uncover the client's reasoning and identify supportive factors. The client reveals that family involvement—specifically exercising with his wife—contributes substantially to his confidence, highlighting the role of social support in sustaining behavior change.

The dialogue demonstrates **goal specificity refinement**, with the coach systematically clarifying ambiguous elements. When the client mentions increasing exercise frequency, the coach probes: "is it walking only, or is there other things that you're doing?" This elicits a more precise formulation incorporating walking, gym sessions, and swimming. The coach further refines the goal by addressing intensity: "in terms of the timing, the intensity, all these also plays, plays a part," prompting the client to describe specific strategies such as walking routes with inclines and treadmill elevation settings.

The interaction also surfaces **peer knowledge integration**, as the client references advice from a friend regarding treadmill inclination for weight loss. The coach validates

this by reframing it in technical language: "you're increasing the intensity to make it from a maybe light intensity to a more moderate intensity." This acknowledgment both affirms the client's existing knowledge and provides professional framing that supports informed decision-making.

Finally, the coach employs **reflective summary** to reinforce progress: "as compared to three months ago where the activity levels were lesser, now you have gym, you have walking, you have swimming. It's really like a progress across the, across time." This technique builds self-efficacy by making improvements explicit and contextualizing the current goal within a trajectory of positive change.

Table 3 maps the dialogue content to the structured coaching script, demonstrating adherence to the goal-setting protocol.

Table 3: Alignment of dialogue with coaching script components for goal-setting sessions.

Script Component	**Evidence in Dialogue**
Ask client to choose personal health goal	Coach references prior session: "you want to walk in the morning three times a day for an hour." Client proposes modification: "if I can increase to three to four times a week."
Explore support, structure, or environments needed	Client identifies family involvement as key enabler: "I can get my wife to join me and we can adjust certain routine." Discusses specific exercise modalities (walking with inclines, gym with treadmill elevation, swimming).
Identify potential obstacles	Implicit in client's conditional phrasing: "if I can, I would... if I can, if I can increase." No major barriers articulated, suggesting readiness for goal.
Refine to SMART behavior goal	Goal progression: *Specific* (walk/gym/swim with increased intensity), *Measurable* (three to four times per week), *Achievable* (confidence level 7), *Relevant* (linked to weight loss objective), *Time-bound* (next month).
Use confidence ruler	Coach applies 1–10 scale implicitly through context from prior goals, then explicitly: "what would your confidence level be?" Client responds with 7. Coach probes reasoning: "How come seven? Not six?"
Improve confidence through strategies	Client confidence already high at 7. Coach focuses on understanding existing strengths rather than incremental improvement. Family support identified as primary confidence factor.
Ask client to restate goals	Coach paraphrases iteratively: "walk or gym lah, three to four times a day." Client confirms and elaborates throughout exchange.
Affirm client's ability	Coach provides explicit affirmation via temporal comparison: "as compared to three months ago where the activity levels were lesser... It's really like a progress across the, across time."