Human Factors in Design, Engineering, and Computing, Vol. 238, 2026, 684–696
https://doi.org/10.54941/ahfe1008300

Human-AI Teaming in Life Coaching: Evaluating LLM Role Fit Through Coaching Competency and Mentor Coach Perspective

Ravneet Kaur and Satarupa Mukherjee

San Francisco Bay University, Fremont, CA 94539, USA

ABSTRACT

Large Language Models (LLMs) are increasingly being used for personal reflection, decision support and coaching conversations. However, their role in life coaching remains unclear, especially when conversations involve sensitive topics such as values, identity, relationships, family expectations, confidence, and major life transitions. This exploratory study examined Human-AI Teaming in life coaching by comparing three LLMs – Claude, Gemini and ChatGPT, to identify their potential Human-AI Teaming roles. Rather than exploring if AI replaces human coaches, the study evaluated if LLMs can support roles like reflection partner, goal-setting assistant, coach-training simulator or mentor coach support aid. The study used six standardized life coaching scenarios and 18 simulated transcripts, evaluated through two lenses – International Coaching Federation (ICF) Core Competencies and a Mentor Coach-informed review. ICF competencies assessed the models' coaching behaviour across 8 competency areas while the Mentor Coach perspective identified their strengths, growth edges, error patterns and risks like giving advice, asking leading questions, over-validation, premature action planning or therapy drift while coaching. Claude achieved the highest observed mean (4.56), followed by Gemini (4.31) and ChatGPT (3.98). Claude demonstrated concise reflection, Gemini structured goal support, and ChatGPT detailed reflection but frequent over-talking. All three were classified as unsuitable for the therapy-adjacent scenario. The analysis identified preliminary role-fit patterns and reinforced the need for human oversight in boundary-sensitive contexts. The study proposes a practical evaluation framework for identifying appropriate LLM role-fit in coaching related Human-AI Teaming. The study follows a human-centered approach where AI is not considered an autonomous life coach but a supportive tool, whose safe and effective use depends on role clarity, professional standards and human oversight.

Keywords: Human-AI teaming, Life coaching, LLMs, Large language models, ICF core competencies, Mentor coaching, AI ethics

INTRODUCTION

Background

LLMs have evolved beyond text completion into interactive systems that can sustain multi-turn conversations. Models like Claude, Gemini and ChatGPT are increasingly used for reflection, decision support, goal setting

Received October 16, 2026; Revised November 22, 2026; Accepted November 29, 2026; Available online December 1, 2026

© 2026 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

and coaching like conversations. Their response may be warm, insightful and confident; however they may not meet the standards expected by Professional Coaching. International Coaching Federation (ICF) defines Professional Coaching as a transformative journey which helps explore new possibilities, and a process that stirs up creativity and fuels growth (International Coaching Federation, 2026). For this study, Life Coaching is defined as a client-centered conversation that helps clarify values and goals, and identify next steps in career, relationships, identity, confidence and life transitions. It is distinct from therapy because it doesn't treat mental health conditions and doesn't take responsibility for client decisions. ICF Core Competencies provide a professional framework for examining ethical practice, coaching mindset, agreements, trust and safety, presence, active listening, evoking awareness, and facilitating client growth (International Coaching Federation, 2026). This study is conducted using Human-AI Teaming. Human-AI Teaming focuses on combining complementary capabilities of human and AI while maintaining role clarity and understanding system limitations. It is important for humans to understand what AI can do, where it's unreliable and which responsibilities should stay with humans (Andrews et al., 2023). In coaching context, human remains the coach, Mentor Coach or instructor while LLM would help with reflection, simulated practice, structured preparation or transcript review.

Research Motivation

This problem matters because people already use LLMs for personal goals, emotional challenges and decisions about identity and relationships. LLMs with their polished language may create trust; however, they may over-validate, direct the client to conclusions, jump to actions before sufficient exploration, over explain or drift into therapy. These behaviours reduce client agency and create safety concerns. Prior research suggests that AI coaching supports goal setting and structured reflection; however when assessed per coaching standards it remains inconsistent and most of the work focuses on workplace, performance, health, or coaching exam contexts (Lai et al., 2025; Passmore, Olafsson, & Tee, 2025; Terblanche et al., 2022). The present study compares multiple LLMs using same life coaching scenario and prompt conditions, assessing them using ICF Core Competencies and an ICF Mentor Coach perspective. The purpose is to identify preliminary strengths, risks and role-fit patterns for different LLMs for Human-AI Teaming in Life Coaching and coach development.

LITERATURE REVIEW

Professional Coaching Quality and Developmental Assessment

ICF Coaching standards provide stronger basis for AI coaching evaluation than helpfulness alone. The ICF Core Competencies include four domains: Foundation, Co-Creating the Relationship, Communicating Effectively, and Cultivating Learning and Growth; and eight competencies: ethical boundaries, coaching mindset, agreements, trust and safety, coaching presence, listening, awareness, and client-owned growth (International Coaching Federation, 2026). They help in distinguishing coaching from advice, consulting, and

therapy. However, competency score alone may not explain what a coach may learn from the conversation. The ICF Mentor Coaching Competencies define mentor coaching as a structured developmental feedback that helps coaches refine their skills and align their practice with professional coaching standards (International Coaching Federation, 2024b). The study uses Mentor Coach perspective as qualitative interpretation of the coaching conversation on top of the competencies' score. This interpretative layer identifies observable strengths, growth edges, and coaching-error patterns such as advice-giving, leading questions, over-validation, premature action planning, therapy drift, and over-talking.

AI Coaching Evidence and Evaluation Approaches

Empirical research shows that AI may be useful in some narrow coaching contexts. In randomized controlled trials, Terblanche et al. (2022) found that AI and human coaches outperformed control conditions for attaining goals. This supports use of AI for goal setting and attainment but it doesn't say anything about emotionally complex and boundary-sensitive life coaching. Passmore, Tee, and Rutschmann (2025) used ICF Competencies and trained ICF assessors to evaluate AI coaching agent in workplace performance and wellbeing conversations. Their work used competency based assessment; however, it left open questions about a comparative study of multiple LLMs, life coaching scenarios and role-fit for Human-AI Teaming. The present study extends this approach, and focuses on how multiple LLMs may fit different roles in supporting life coaching and coach development.

Pavlovic et al. (2024) compared humans and multiple LLMs on ICF-mimicking examination tasks. They found that LLMs perform well on structured coaching judgment tasks but struggle with ambiguity and context. This raises concerns about their ability to maintain presence, preserve client agency, and manage emotional and ethical complexity throughout coaching conversations.

Systematic and scoping reviews point to methodological fragmentation. Lai et al. (2025) report inconsistent evaluation strategies and limited real-world validation of LLM based health and exercise coaching. Though health and exercise coaching are different from life coaching, these reviews reveal a broader issue: AI-supported coaching systems are being developed much faster than consistent methods to evaluate quality, safety and appropriate use of these systems. This supports need of a structured framework rather than just claiming if AI coaching "works" or "does not work".

Human-AI Teaming, Hybrid Coaching, and Ethical Boundaries

Human-AI Teaming is considering AI to provide support rather than AI system replacing the coach. Andrews et al. (2023) discuss that shared mental models help in effective teaming when capabilities, limitations and task roles are clear. Effective Human-AI Teaming also depends on the amount of trust on AI system. Users must understand what an AI system can contribute, and when the AI outputs need to be verified, interpreted and intervened by humans.

Loughnane et al. (2025) found that human, AI and hybrid approaches are feasible and acceptable in digital health coaching while hybrid approaches still need to be refined with human depth and judgement when scaling. Applied to life coaching, these findings mean that the human will remain the coach, mentor coach, evaluator, instructor, or learner while AI supports reflection, simulation, preparation, structured practice, or transcript review. Therefore, ICF Competency Framework and Mentor Coach perspective help to determine the quality and limits of these possible AI roles. This connection also places the study in the area of human-computer interaction by examining how humans understand, trust, use and constrain AI in sensitive coaching contexts.

Ethical boundaries become important in case coaching conversations start approaching mental health concerns. ICF AI Coaching Framework and Standards emphasize privacy, bias reduction, transparency, coaching-science alignment, and clear difference between coaching and therapy. These concerns are relevant for life coaching because the client may start with a standard coaching conversation like a reflection or goal-setting but may lead to emotional vulnerable experiences.

Scholich et al. (2025) found that standard chatbots affirm and reassure better than the human therapists but were unsuitable as a therapeutic agent in crisis situations. This evidence shows that apparent empathy should not be considered as a proof of safe and appropriate support. A model may sound caring but may give excessive advice, over-validate, talk too much, ask leading questions or miss to direct the client to a qualified human when the need arise. This is relevant to the study's therapy-adjacent scenario that supports evaluating boundary recognition and client agency and appropriateness of proposed Human-AI Teaming role.

Research Gaps

The literature establishes that AI coaching maybe useful, ICF competencies can support evaluation of coaching scenarios and that ethical boundaries and role clarity matters. However, there are still some clear gaps. Few studies compare general-purpose LLMs in the same complete coaching conversations. The current study combines ICF competencies evaluation with Mentor Coach perspective, explicit coaching error coding, and Human-AI Teaming role-fit mapping. These combinations are rarely used in the previous studies. Life coaching scenarios involving values, identity, relationships, family expectations, client agency, and therapy-adjacent distress are also less represented in earlier studies. The present study does a standardized 3 × 6 comparison of three Coach Agents (Claude, Gemini and ChatGPT) across six life coaching scenarios. Its main contribution is a pilot evaluation framework that connects professional coaching quality, developmental risk analysis, and role-specific Human-AI Teaming recommendations. These possible roles include reflection partner, structured goal-setting assistant, coach-training simulator, Mentor Coach support aid, and not suitable for boundary-sensitive contexts.

Research Questions

The study addressed four exploratory research questions:

RQ1: How do 3 models, Claude, Gemini, and ChatGPT, perform across ICF Core Competencies when responding to standardized life coaching scenarios?
RQ2: What strengths, growth edges, and coaching-error patterns emerge when the same responses are evaluated from an ICF Mentor Coach perspective?
RQ3: What preliminary Human-AI Teaming role-fit patterns emerge for each LLM in life coaching and coach development?
RQ4: How do the models differ in ethically sensitive situations, such as advice-seeking, emotional distress, cultural/family conflict, and therapy-adjacent concerns?

METHODOLOGY

Research Methodology

The study used a convergent mixed-methods methodology within an exploratory comparative pilot. The quantitative strand consisted of descriptive transcript-level scores across eight ICF Core Competencies. The qualitative strand consisted of Mentor Coach review of strengths, growth edges, coaching-error patterns, and role fit. Both strands examined the same transcripts and were integrated.

Research Design

The study used a repeated-scenario comparative design with a 3 × 6 factorial structure: three Coach Agents (Claude, Gemini, and ChatGPT) and six life coaching scenarios, including boundary-sensitive situations. Each model completed each scenario once, generating 18 transcripts. Grok 4 served as a fixed Client Agent to maintain consistency in client behaviour across conditions.

Table 1: Variables and controls.

Design Element	Operationalization	Purpose
Independent variable	Coach Agent: Claude Sonnet 5, Gemini 3.5 Flash, ChatGPT 5.5	Compare observed coaching behaviour by model.
Scenario factor	Six standardized life-coaching scenarios	Test performance across structured, complex, and sensitive contexts.
Controlled conditions	Same Client Agent, scenario instructions, Coach Agent prompt, transcript-completion rule, and rubric	Increase within-study comparability.

(Continued)

Table 1: Continued.

Design Element	Operationalization	Purpose
Dependent measures	Eight ICF scores, average score, six binary error codes, qualitative notes, and role-fit label	Capture both coaching quality and developmental risk.
Unit of analysis	One complete simulated coaching transcript	Evaluate sustained behaviour across the conversation.

Primary threats to validity included LLM nondeterminism, the use of an LLM as the Client Agent, one generated transcript per condition, a single evaluator, and model-version changes. These threats were not ignored; they defined the interpretive boundary of the pilot. The results were therefore reported as observed sample patterns and framework-testing evidence rather than stable estimates of model performance.

EXPERIMENTAL DESIGN

Data

The final dataset contained 18 simulated life-coaching transcripts collected on July 7, 2026. Grok 4 acted as the fixed Client Agent. Claude Sonnet 5, Gemini 3.5 Flash, and ChatGPT 5.5 acted as Coach Agents. Each Coach Agent completed the same six scenarios: goal setting and procrastination; values conflict involving family expectations and personal dreams; relationship boundaries and resentment; life transition and identity uncertainty; advice seeking and client agency; and therapy-adjacent emotional distress. The raw transcript record was preserved unchanged. A separate cleaned spreadsheet stored one row per transcript and included transcript ID, scenario, model, date, completeness status, eight competency scores, average score, six error codes, Mentor Coach-informed notes, role-fit category, and rationale. The dataset was created for this research and is not licensed as a public corpus. It contains generated text rather than human-subject data. Ethical protections were built into the design by avoiding real client material, retaining boundary responses as data, and evaluating therapy-adjacent prompts for recognition and redirection rather than for therapeutic effectiveness.

Table 2: Dataset summary.

Item	Description
Client Agent	Grok 4
Coach Agents	Claude Sonnet 5; Gemini 3.5 Flash; ChatGPT 5.5
Scenarios	Six standardized life-coaching scenarios
Total transcripts	18
Collection date	July 7, 2026
Unit of analysis	Complete simulated coaching transcript
Analysis	Descriptive ICF scoring plus Mentor Coach qualitative review

Test Plan

Each transcript was scored from 1 to 5 on eight ICF Core Competencies. A score of 5 represented consistent and strong evidence; 4, strong but inconsistent evidence; 3, acceptable but uneven evidence; 2, limited evidence; and 1, weak, absent, unsafe, or role-inappropriate evidence. Transcript average was arithmetic mean of the eight scores. Model-level and scenario-level averages were calculated.

Six coaching-error patterns were coded at the transcript level using a binary rule (0 = absent, 1 = present). The patterns were advice-giving, leading questions, over-validation, premature action planning, therapy drift, and over-talking or lecture style. Frequencies represent counts of transcripts containing the pattern and are not severity ratings. The qualitative review recorded strengths, growth edges, evidence supporting or weakening competency scores, and the most appropriate preliminary role-fit category: reflection partner, structured goal-setting assistant, coach-training simulator, mentor coach support aid, or not suitable for sensitive contexts.

Role-fit classifications were based on an integrated review of ICF competency scores, observed coaching-error patterns, and Mentor Coach qualitative judgments. Each transcript was assigned the role that best matched its observed strengths, limitations, and ethical boundaries.

No train-validation-test split was applicable because the project did not train a predictive model. No inferential statistical test was conducted because one transcript per condition does not provide a within-cell estimate of LLM response variance. Meaningful-difference thresholds were not defined before data collection, so no post hoc threshold was used to label observed score differences as statistically significant. The analysis is descriptive and exploratory.

Design of Experiments

The experiment followed the procedure below:

1. Create six standardized scenarios that span ordinary goal work, values and identity concerns, interpersonal boundaries, advice pressure, and therapy-adjacent distress.
2. Provide Grok with the Client Agent prompt and the relevant scenario persona.
3. Provide each Coach Agent with the same ICF-aligned coaching prompt.
4. Transfer messages between the Client Agent and Coach Agent until the Client Agent ended the simulation and the Coach Agent supplied its final response.
5. Save the complete transcript with a consistent transcript ID, date, scenario, model names, and prompt information.
6. Check completeness and preserve the raw transcript before scoring or coding.
7. Apply the same ICF rubric and Mentor Coach feedback to all 18 transcripts.
8. Integrate scores, error frequencies, qualitative patterns, and role-fit labels to answer the research questions.

FINDINGS

Conduct of Experiments

The proposal allowed for additional models if time permitted, but no additional models were included. The proposal also identified multiple evaluators as desirable; the final scoring and qualitative review were completed by one researcher with ICF PCC-level coaching and Mentor Coach experience. The evaluator's domain expertise supported consistent application of coaching concepts, while the absence of a second rater limited reliability assessment.

Experimental Result

Table 3: Observed average ICF scores by coach agent: overall and foundational competencies.

Coach Agent	Overall	Ethical Practice	Coaching Mindset	Agreements	Trust & Safety
Claude	4.56	4.83	4.83	4.00	4.83
Gemini	4.31	4.83	4.00	4.00	4.83
ChatGPT	3.98	4.83	3.83	3.83	4.67

Table 4: Observed average ICF scores by coach agent: communication and growth competencies.

Coach Agent	Presence	Active Listening	Evokes Awareness	Client Growth
Claude	4.67	4.83	4.67	3.83
Gemini	4.50	4.33	4.17	3.83
ChatGPT	2.67	4.17	4.17	3.67

Table 5: Observed average ICF score by scenario.

Scenario	Average ICF Score
S1 Goal setting and procrastination	4.29
S2 Values conflict	4.46
S3 Relationship boundaries	4.54
S4 Life transition	4.42
S5 Advice seeking and client agency	4.29
S6 Therapy-adjacent emotional distress	3.71

Table 6: Coaching error-pattern frequency by coach agent.

Coaching Error Pattern	Claude	Gemini	ChatGPT
Advice-giving	0	0	0
Leading questions	0	1	1
Over-validation	0	5	0
Premature action planning	0	3	0

(Continued)

Table 6: Continued.

Coaching Error Pattern	Claude	Gemini	ChatGPT
Therapy drift	1	0	1
Over-talking / lecture style	0	0	6
Total Error Pattern	1	9	8

Note. Values are transcript frequency counts out of six transcripts per Coach Agent; they are not severity ratings.

Table 7: Human-AI teaming role-fit frequency by coach agent.

Role-Fit Category	Claude	Gemini	ChatGPT
Reflection partner	3	2	4
Structured goal-setting assistant	1	3	0
Coach-training simulator	1	0	0
Mentor Coach support aid	0	0	1
Not suitable for sensitive contexts	1	1	1

Note. Role-fit values show the number of scenario transcripts assigned to each category. The labels are exploratory judgments derived from integrated score and transcript review.

ANALYSIS

RQ1: Observed ICF Core Competency Performance

Claude had the highest observed overall mean at 4.56, followed by Gemini at 4.31 and ChatGPT at 3.98. These differences were not tested for statistical significance and should not be interpreted as stable population-level effects. All three models scored high for Ethical Practice. Claude showed consistently high observed scores in Coaching Mindset, Presence, Active Listening, and Evokes Awareness. Gemini showed relatively strong performance in Trust, Safety and Presence, with lower observed scores in Coaching Mindset and Evokes Awareness. ChatGPT demonstrated strong reflection, but its lower Presence score of 2.67 led to its lower overall average. Overall, the results suggest observable differences in how the three models demonstrated ICF-aligned coaching behaviors within the 18 transcripts.

RQ2: Strengths, Growth Edges, and Coaching-Error Patterns

The Mentor Coach review identified differences not fully visible in the average competency scores. Claude's main observed strength was concise and reflective interaction that generally preserved client ownership. Its primary growth edge involved boundary handling, with therapy drift in one transcript. Gemini converted broad concerns into structured reflection and manageable next steps. However, over-validation appeared in five transcripts, premature action planning in three, and a leading question in one. ChatGPT generated detailed reflection, but over-talking or lecture style appeared in all six transcripts.

This pattern reduced its observed Presence score and shifted the interaction from coaching dialogue toward explanation. These findings indicate that competency scores alone do not fully capture coaching quality. The qualitative review revealed specific developmental patterns, that are important when assessing whether an LLM models appropriate coaching behavior.

RQ3: Preliminary Human-AI Teaming Role-Fit Patterns

The integrated analysis suggested different preliminary Human-AI Teaming roles for the three models. Claude was mostly categorized as a reflection partner and once as a coach-training simulator because of its concise responses, reflective questioning, and preservation of client agency. Gemini was mostly categorized as a structured goal-setting assistant because of its tendency to organize concerns, identify manageable steps, and support forward movement. However, its patterns of over-validation and premature action planning suggest that human oversight would still be necessary. ChatGPT was mostly categorized as a reflection partner and appeared once as a mentor coach support aid. Its detailed responses provided substantial material for analysis and transcript review, although their length made them less suitable for direct coaching dialogue without response controls.

RQ4: Differences in Ethically Sensitive Situations

The models showed their greatest limitations in the therapy-adjacent emotional-distress scenario, where the lowest observed average at 3.71. All three models were categorized as not suitable for sensitive contexts. Claude and ChatGPT each received one therapy-drift code. Gemini's response was evaluated as safer while also being less developmentally useful for coaching growth. In advice-seeking scenarios, the models generally avoided direct advice-giving but sometimes rushed toward action or subtly directed the client. In values-conflict and family-expectation scenarios, the models reflected competing priorities, although depth, neutrality, and client ownership varied. These findings suggest that safety cannot be inferred from warmth, fluency, or a high average score alone. Sensitive conversations require examination of boundaries, client agency, response content, and the model's ability to recognize when human support is needed.

DISCUSSION

Interpretation in Relation to Prior Research

The findings reinforce AI coaching as a collection of bounded capabilities rather than "if an AI coach works". The observed usefulness of structured goal support is consistent with Terblanche et al. (2022), who found that AI coaching could support goal attainment. Gemini as a structured goal-setting assistant illustrates this. Building on Passmore, Tee, and Rutschmann (2025), the present study used professional coaching competencies to evaluate multiple LLMs along with a Mentor Coach perspective. This second lens was consequential: even high-scoring models had weaknesses, such as over-talking, over-validation, or premature action planning. The results support

Pavlovic et al. (2024), showing models performed well on structured coaching conversations, yet the therapy-adjacent scenario required boundaries and role awareness. The results also align with Scholich et al. (2025), who suggested that supportive responses alone do not make AI suitable for therapy. From a Human-AI Teaming perspective, the role-fit analysis supports the importance of shared mental models and appropriate trust (Andrews et al., 2023). A model may be useful for one task and poorly suited to another. This is also consistent with Loughnane et al. (2025), who argue that hybrid coaching approach requires refinement to combine AI scalability with human depth and judgment.

Practical and Ethical Implications

The value of the study lies in role clarity. An LLM used as a reflection partner can help a person articulate thoughts, while a structured goal-setting assistant can help organize next steps. In coach development, an LLM may support simulated practice or help for mentor coach review. These roles require explicit instructions, response-length controls, privacy protections, and human ethical oversight. The findings do not support autonomous AI use in sensitive contexts. Every model was categorized as unsuitable for therapy-adjacent scenario. This aligns with the ICF AI Coaching Framework and Standards, which emphasizes safeguards, transparency and boundaries between coaching and therapy (International Coaching Federation, 2024a). Human judgment remains central in conversations involving emotional distress, safety risk or decisions beyond coaching scope.

Design Boundaries and Interpretive Limits

Several study boundaries were intentional choices to support controlled pilot.

- Simulated scenarios were selected to protect client privacy and to allow consistent comparisons across models.
- A fixed LLM Client Agent reduced variation but may have influenced results through its interaction with each Coach Agent.
- One transcript was generated for each model-scenario cell. This was sufficient to pilot the rubric, coding process, and role-fit framework across 18 complete conversations, but it did not measure run-to-run variation.
- One domain-expert evaluator applied the rubric and Mentor Coach review. This supported consistent coaching interpretation, but may have introduced potential evaluator bias.
- The sample supported only descriptive statistics, without predefined significance thresholds. Model versions and hosted-system behaviour may also change over time. The findings are bounded to tested prompts, scenarios, model versions, date, and evaluator. Nevertheless, the study contributes to framework development and preliminary pattern identification.
- Alternative explanations should be considered. ChatGPT's over-talking pattern, Gemini's structured action orientation or Claude's concision require repeated trials with controlled generation parameters to separate stable model tendencies from individual conversational draws.

CONCLUSION

This exploratory pilot compared Claude, Gemini, and ChatGPT as life-coaching Coach Agents across six standardized simulated scenarios. Eighteen complete transcripts were evaluated using a two-layer framework: descriptive scoring across eight ICF Core Competency areas and Mentor Coach qualitative review of strengths, growth edges, coaching-error patterns, and Human-AI Teaming role fit. In the observed sample, Claude had the highest average score, Gemini showed a strong structured goal-support pattern, and ChatGPT produced rich reflection but frequently over-talked. The therapy-adjacent scenario was the weakest across models and required human oversight. The main contribution is a practical evaluation framework that connects ICF coaching competencies with developmental risk analysis and role-specific Human-AI Teaming. The findings show how the same model can be useful for one bounded task and inappropriate for another. LLMs may support reflection, preparation, simulation, and structured practice, but they should not replace human coaches or be treated as appropriate for therapy-adjacent and emotionally sensitive situations.

RECOMMENDATION

Based on the observed results, the following are the recommendations:

- Use LLMs as bounded support tools rather than autonomous coaches, with the human retaining responsibility for ethics, interpretation, and accountability.
- Match the tool to the task: reflection support, structured goal clarification, simulated practice, or transcript-review assistance.
- Use explicit coaching-boundary prompts, privacy protections, and response-length controls to reduce advice drift, over-talking, and role confusion.
- Do not use the tested systems as substitutes for mental health professionals or as independent support in crisis or therapy-adjacent contexts.
- Treat model-specific role-fit findings as preliminary and time-bounded because model behavior can change across versions and runs.

Future research should:

1. Repeat each model-scenario condition at least three times, producing a minimum of 54 transcripts, and report the mean, range, and within-cell variation.
2. Predefine practical-difference thresholds before data collection, such as a model-level mean difference of at least 1.0 point on the 5-point ICF scale, then select statistical methods appropriate to the repeated design.
3. Use multiple qualified evaluators, calculate inter-rater reliability, and document how disagreements in scores and role-fit labels are resolved.
4. Control and report model version, system prompt, temperature or generation settings where available, interface, and collection sequence to improve reproducibility.

5. Extend the framework to real or semi-real client interactions with informed consent, privacy safeguards, risk protocols, and clear separation from therapy.

REFERENCES

Andrews, R. W., Lilly, J. M., Srivastava, D., & Feigh, K. M. (2023). The role of shared mental models in human-AI teams: A theoretical review. *Theoretical Issues in Ergonomics Science, 24*(2), 129–175. https://doi.org/10.1080/1463922X.2022.2061080

International Coaching Federation. (2024a, November 8). *Artificial intelligence coaching framework and standards*. https://coachingfederation.org/resource/icf-artificial-intelligence-ai-coaching-framework-and-standards/

International Coaching Federation. (2024b, November 13). *ICF mentor coaching competencies*. https://coachingfederation.org/resource/mentor-coaching-competencies/

International Coaching Federation. (2026, April 14). *ICF core competencies*. https://coachingfederation.org/resource/icf-core-competencies/

Lai, X., Lai, Y., Chen, J., Huang, S., Gao, Q., & Huang, C. (2025). Evaluation strategies for large language model-based models in exercise and health coaching: Scoping review. *Journal of Medical Internet Research, 27*, e79217. https://doi.org/10.2196/79217

Loughnane, C., Laiti, J., O'Donovan, R., & Dunne, P. J. (2025). Systematic review exploring human, AI, and hybrid health coaching in digital health interventions: Trends, engagement, and lifestyle outcomes. *Frontiers in Digital Health, 7*, Article 1536416. https://doi.org/10.3389/fdgth.2025.1536416

Passmore, J., Olafsson, B., & Tee, D. (2025). A systematic literature review of artificial intelligence (AI) in coaching: Insights for future research and product development. *Journal of Work-Applied Management*. https://doi.org/10.1108/JWAM-11-2024-0164

Passmore, J., Tee, D., & Rutschmann, R. (2025). "Getting better all the time": Using professional human coach competencies to evaluate the quality of AI coaching agent performance. *Coaching: An International Journal of Theory, Research and Practice*, 1–17. https://doi.org/10.1080/17521882.2025.2510643

Pavlovic, J., Krstic, J., Mitrovic, L., Babic, D., Milosavljevic, A., Nikolic, M., Karaklic, T., & Mitrovic, T. (2024). *Generative AI as a metacognitive agent: A comparative mixed-method study with human participants on ICF-mimicking exam performance*. arXiv. https://arxiv.org/abs/2405.05285

Scholich, T., Barr, M., Wiltsey Stirman, S., & Raj, S. (2025). A comparison of responses from human therapists and large language model-based chatbots to assess therapeutic communication: Mixed methods study. *JMIR Mental Health, 12*, e69709. https://doi.org/10.2196/69709

Terblanche, N., Molyn, J., de Haan, E., & Nilsson, V. O. (2022). Comparing artificial intelligence and human coaching goal attainment efficacy. *PLOS ONE, 17*(6), e0270255. https://doi.org/10.1371/journal.pone.0270255