



King's Research Portal

DOI:

[10.1016/j.artmed.2024.103004](https://doi.org/10.1016/j.artmed.2024.103004)

Document Version

Peer reviewed version

[Link to publication record in King's Research Portal](#)

Citation for published version (APA):

Ong, Q. C., Ang, C.-S., Chee, D. Z. Y., Lawate, A., Sundram, F., Dalakoti, M., Pasalic, L., To, D., Erlikh fox, T., Bojic, I., & Car, J. (2024). Advancing health coaching: A comparative study of large language model and health coaches. *Artificial Intelligence in Medicine*, 157, Article 103004. <https://doi.org/10.1016/j.artmed.2024.103004>

Citing this paper

Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights

Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the Research Portal

Take down policy

If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Advancing Health Coaching: A comparative study of Large Language Model and Health Coaches

Qi Chwen Ong ^{a,b,1}, Chin-Siang Ang ^a, Davidson Zun Yin Chee ^a, Ashwini Lawate ^a, Frederick Sundram ^c, Mayank Dalakoti ^{d,e}, Leonardo Pasalic ^{f,g}, Daniel To ^h, Tatiana Erlikh Fox ^{a,i}, Iva Bojic ^a, Josip Car ^{a,j}

^a Lee Kong Chian School of Medicine, Nanyang Technological University, 11 Mandalay Rd, 308232, Singapore

^b School of Public Health, Imperial College London, 90 Wood Ln, London W12 0BZ, United Kingdom

^c Department of Psychological Medicine, Faculty of Medical and Health Sciences, University of Auckland, Auckland 1023, New Zealand

^d Department of Cardiology, National University Heart Centre, 5 Lower Kent Ridge Rd, 119074 Singapore

^e Cardiovascular Metabolic Disease Translational Research Program, National University of Singapore, Singapore

^f Haematology, Sydney Centres for Thrombosis and Haemostasis, Institute of Clinical Pathology and Medical Research (ICPMR), NSW Health Pathology, Westmead Hospital, Westmead, NSW Australia

^g Westmead Clinical School, University of Sydney, Westmead, NSW Australia

^h Department of Medicine, University of Wisconsin-Madison, Madison, WI, United States

ⁱ Onze Lieve Vrouwen Gasthuis, Jan Tooropstraat 164, 1061 AE Amsterdam, Netherlands

^j School of Life Course & Population Sciences, King's College London, Strand, WC2R 2LS, United Kingdom

¹ Corresponding author. Email: qichwen.ong@ntu.edu.sg

Abstract

Objective:

Recent advances in large language models (LLM) offer opportunities to automate health coaching. With zero-shot learning ability, LLMs could revolutionize health coaching by providing better accessibility, scalability, and customization. The aim of this study is to compare the quality of responses to clients' sleep-related questions provided by health coaches and an LLM.

Design, Setting, and Participants:

From a de-identified dataset of coaching conversations from a pilot randomized controlled trial, we extracted 100 question-answer pairs comprising client questions and corresponding health coach responses. These questions were entered into a retrieval-augmented generation (RAG)-enabled open-source LLM (LLaMa-2-7b-chat) to generate LLM responses. Out of 100 question-answer pairs, 90 were taken out and assigned to three groups of evaluators: experts, lay-users, and GPT-4. Each group conducted two evaluation tasks: (Task 1) a single-response quality assessment spanning five criteria—accuracy,

readability, helpfulness, empathy, and likelihood of harm—rated on a five-point Likert scale, and (Task 2) a pairwise comparison to choose the superior response between pairs. A suite of inferential statistical methods, including the paired and independent sample t-tests, Pearson correlation, and chi-square tests, were utilized to answer the study objective. Recognizing potential biases in human judgment, the remaining 10 question-answer pairs were used to assess inter-evaluator reliability among the human evaluators, quantified using the interclass correlation coefficient and percentage agreement metrics.

Results:

Upon exclusion of incomplete data, the analysis included 178 single-response evaluations (Task 1) and 83 pairwise comparisons (Task 2). Expert and GPT-4 assessments revealed no discernible disparities in health coach and LLM responses across the five metrics. Contrarily, lay-users deemed LLM responses significantly more helpful than that of human coaches ($p < 0.05$). LLM responses were preferred in the majority (62.25%, $n=155$) of the aggregate 249 assessments, with all three evaluator groups favoring LLM over health coach inputs. While GPT-4 rated both health coach and LLM responses significantly higher than experts in terms of readability, helpfulness, and empathy, its ratings on accuracy and likelihood of harm aligned with those of experts. Response length positively correlated with accuracy and empathy scores, but negatively affected readability across all evaluator groups. Expert and lay-user evaluators demonstrated moderate to high inter-evaluator reliability.

Conclusion:

Our study showed encouraging findings by demonstrating that RAG-enabled LLM has comparable performance to health coaches in the domain tested. Serving as an initial step towards the creation of more sophisticated, adaptive, round-the-clock automated health coaching systems, our findings call for more extensive evaluation which could assist in the development of the model that could in the future lead to potential clinical implementation.

Keywords: Retrieval-augmented generation, AI health coaching, Human evaluation, Q/A system, Sleep

1.0 Introduction

Non-communicable diseases are linked to lifestyle, with a significant portion of associated risks attributable to behavioral factors.¹ The traditional doctor-centric healthcare model has shown limitations in effectively fostering health behavior change.² In contrast, health coaching has emerged as a patient-centered strategy, promoting positive changes in lifestyle habits such as nutrition, physical activity, and weight management,

improving medication adherence, assisting with better glycemic control in patients with chronic diseases, and improving patients' physiological, behavioral, psychological and social outcomes.³⁻⁶ It is defined as “the practice of health education and health promotion within a coaching context, to enhance the wellbeing of individuals and to facilitate the achievement of their health-related goals”.⁷ Health coaching interventions are becoming more common in health management programs across various settings, from physician offices to population-wide initiatives.⁸ The durability of clinical benefits post-intervention suggests that the behavioral changes facilitated by health coaching have the potential for long-term sustainability.⁹

While traditional health coaching methods have proven effective, they often hinge on direct human interaction, which may impede scalability and customization.¹⁰ The emergence of advanced natural language processing (NLP) tools, exemplified by OpenAI's ChatGPT, offers a unique opportunity to overcome these barriers. The burgeoning domain of health coaching stands at a critical crossroad, poised to significantly elevate the sustainability and reach of health guidance by integrating cutting-edge artificial intelligence (AI) technologies.

The use of LLM in health coaching is motivated by the intrinsic constraints of existing automated health coaching platforms, which largely depend on scripted interactions.¹¹ Although these systems are beneficial, they frequently fall short in delivering personalized, adaptive, and nuanced counsel, essential for effective behavioral change. The quintessence of successful health coaching resides in its adaptability to an individual's distinct situation, needs, and predilections, providing customized recommendations that strike a chord on a personal level. Conventional automated systems, with their standardized and rule-based dialogues, do not meet this standard, dispensing generic advice that may lack relevance or motivational value for diverse users.

LLM, with its zero-shot learning capabilities, marks a significant departure from its predecessors. It generates responses in real-time, dynamically crafted from its extensive training data, rather than relying on pre-scripted answers. This innovation introduces a degree of personalization and adaptability that was once beyond the reach of automated systems in health coaching. The present study compares the responses from human health coaches with those produced by LLM to assess the practicality and efficacy of integrating such AI technologies into health coaching. The objective is not to supplant human coaches but to supplement and refine the health coaching process, thereby enhancing its accessibility, scalability, and customization.

2.0 Materials and methods

We conducted a comparative analysis to examine the quality of responses from health coaches and a retrieval-augmented generation (RAG)-enabled LLM. Employing a tripartite evaluation methodology—expert (physicians), lay-user (clients), and LLM-based (GPT-4) assessments—we gauged the quality of these responses. Health coaches were university students trained in sleep coaching, while the LLM used was Llama-2–7b-chat model. Each group of evaluators (i.e., expert, lay-user, and GPT-4) conducted two evaluation tasks: (Task 1) a single-response quality assessment spanning five criteria, and (Task 2) a pairwise comparison to express their preference between health coach and LLM responses.

2.1 Collection of health coach responses

Responses were amassed from transcripts of 120 coaching dialogues, part of a pilot randomized controlled trial (RCT) spanning March to April 2023. Participants of the pilot RCT, drawn from a university, engaged in remote health coaching intervention on sleep via text messaging, facilitated by health coaches who were supported by a domain-specific question-answering system (QA) system.¹² This system was designed to provide evidence-based answers to sleep-related questions. Health coaches were university undergraduate and postgraduate students who completed a 7-week health coaching training course offered by the University’s Centre for Population Health Sciences and passed the assessments. Before the study started, the health coaches had limited real-world health coaching experience. During the study, they were given the option to utilize, modify, or disregard the answers generated by the QA system. The conversations between participants and health coaches were recorded on Microsoft Teams (version 2023033810). Informed consent for the use of these recordings in future research was obtained from all parties involved. Ethics approval was granted by the Institutional Review Board (IRB) of Nanyang Technological University Singapore (reference number: IRB-2021-739).

We curated all recorded conversations to create a dataset of question-response pairs (i.e., a unique client question and corresponding health coach response). To ensure privacy, all personal identifiers were expunged. Question-response pairs for which de-identification was not possible without changing the original content were removed from the dataset. We randomly selected 100 question-response pairs from a pool of 126 as study evaluation points.

2.2 Collection of LLM responses

We used the Llama-2–7b-chat model,¹³ an open-source LLM released by Meta AI,¹⁴ in combination with the LangChain framework.¹⁵ We chose open-source LLM for its affordability, flexibility, and option for optimization and customization. Llama 2 was one of the newest and most accessible open-source LLMs at

the time of this study. The 7-billion parameter Llama 2 model was selected as it was faster and required fewer computational resources compared to the 70-billion and 13-billion parameter models. We used the default parameters for this model, except for the following adjustments: the maximum number of new tokens was set to 512 and the temperature was set to 0.5. To demonstrate the feasibility of this approach in low-resource settings, we utilized a setup on a local laptop with i7-10510U CPU and 16.0 GB RAM. Prompts from the LangChain framework enable users to create dynamic templates that support generating more precise and relevant responses to the user's questions. In our application, we designed a custom prompt (**Supplement Note 1**) that uses the user's question and the context behind it to prevent the generation of irrelevant responses. Through retrieval augmented generation (RAG),¹⁶ external data from the 'SleepQA' dataset were fetched and fed to the LLM to generate the response.¹² We used the 'SleepQA' dataset comprising more than one thousand medically reviewed articles on sleep health to address the sleep-related questions. Moreover, we used the *RecursiveCharacterTextSplitter* to split large documents from the 'SleepQA' dataset into smaller parts,¹⁷ *sentence-transformers/all-MiniLM-L6-v2* model from the *HuggingFaceEmbeddings* library in LangChain to create document embeddings, and the Facebook AI similarity search vector to store them.¹⁸ After the data had been stored in the vector store database, the *RetrievalQA* algorithm within LangChain was used to fetch the most relevant information. Finally, both the user's query and the information obtained through the *RetrievalQA* algorithm were inputted into the LLM model to generate a response. Since health coaches generally provided only one response to each client's question, the prompt for Llama-2-7b-chat model was run once only per question. Re-running the prompts for multiple outputs, while ensuring reproducibility, deviates from real-world coaching practices and create an unfair comparison.

2.3 Human evaluation design

Although automatic metrics have gained popularity for their cost, speed and repeatability, they are difficult to interpret and do not correlate with human evaluation when used for approximating text quality.¹⁹ Human evaluation remains the gold standard for general assessment of overall NLP system quality.¹⁹ To evaluate the responses comprehensively, we employed both expert-focused (physicians) and reader-focused designs (clients) in human evaluation as they complement one another.²⁰

Expert evaluations involved a group of five licensed physicians practicing in New Zealand, Australia, Singapore, the United States, and the Netherlands. They have backgrounds in different medical disciplines, including general internal medicine, psychiatry, hematology, and cardiology. Physicians were chosen as expert evaluators based on the assumption that they have a deep knowledge of sleep physiology and sleep

health recommendations, even though they did not have substantial health coaching experience. Five university students were involved in lay-user evaluations. University students were selected as lay users as they are the target population of the health coaching on sleep intervention. Evaluators were not involved in the collection of study samples or in any way exposed to the dataset before the evaluation process.

2.4 Human evaluation tasks

Our evaluation process involved two tasks (see **Figure 1**): Single-response grading (i.e., Task 1) and pairwise comparison (i.e., Task 2). For single-response grading, evaluators assessed the quality of responses according to pre-defined criteria (see **Table 2**). Each question came with a corresponding response from either a health coach or LLM. Each group of evaluators was asked to perform a total of 180 unique evaluation points (i.e., 90 responses from the health coach and 90 responses from LLM). For pairwise comparison, evaluators were presented with a question and two responses (i.e., one from a health coach and one from LLM) and asked to choose the better response of the two. Each group of evaluators completed a total of 90 unique evaluation points. Evaluators were blinded to the source of responses in both tasks. Furthermore, we asked all individual evaluators to complete the same set of additional 10 evaluation points for both tasks to measure inter-evaluator agreement.

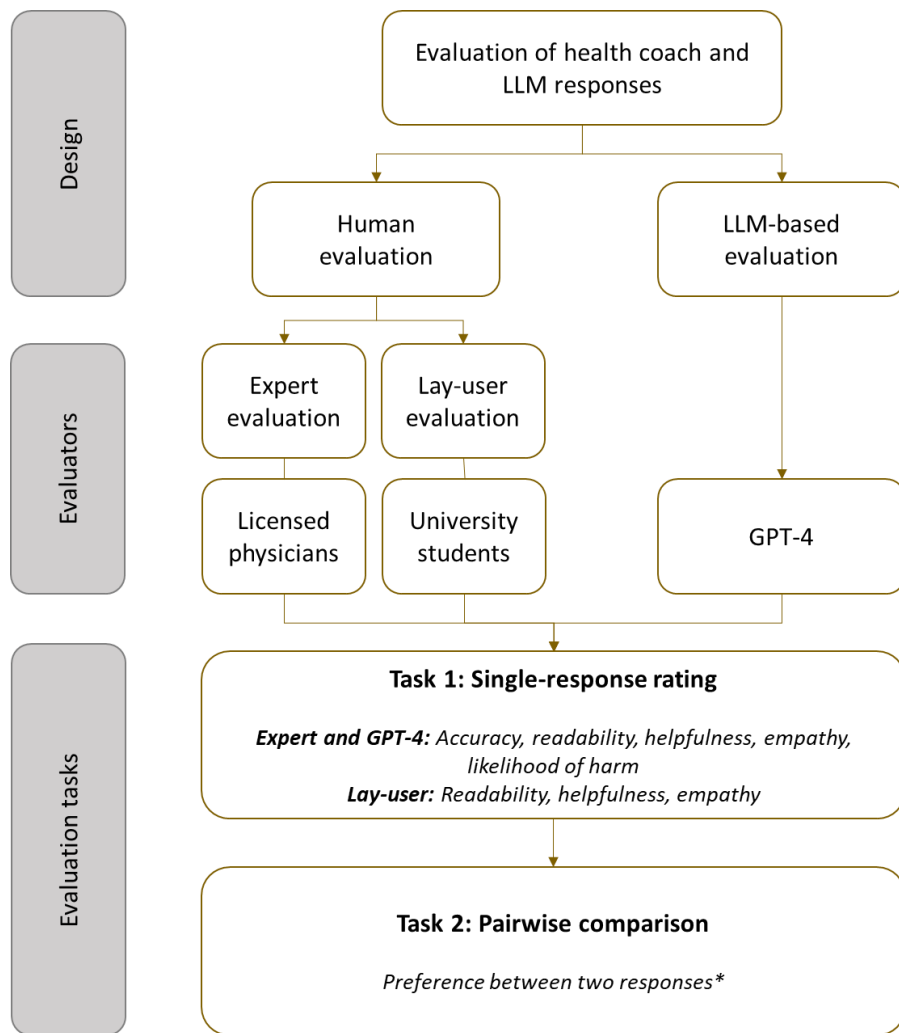


Figure 1. Human evaluation of health coach and LLM responses.

2.5 Evaluation criteria

As there is little consensus on the human evaluation framework in NLP, we based our metrics on the pilot framework developed by Singhal et al. and a scoping review of human evaluation in Natural Language Processing (NLG).^{21,22} Given that content-based errors such as factual inaccuracies are more common than fluency-based errors in large language models,²³ we omitted fluency-related aspects of quality in this study. For the first task, expert evaluators were asked to assess each response for its accuracy, readability, helpfulness, empathy, and likelihood of harm, while lay-user evaluators judged its readability, helpfulness, and empathy (see Figure 1). Details on the quality criteria are presented in Table 1.

Table 1. Quality criteria used for evaluation of health coach and LLM responses.

Quality Criterion	Definition	5-point Likert Scale
-------------------	------------	----------------------

Accuracy	Accuracy refers to the degree to which the information provided aligns with established medical knowledge, clinical guidelines, and evidence-based practices. An accurate response is medically sound and free from factual errors, misinformation, or misinterpretation of medical data.	1: Completely inaccurate 2: Mostly inaccurate 3: Somewhat accurate 4: Mostly accurate 5: Completely accurate
Readability	Readability refers to the ease of understanding (use of plain language without excessive jargon or complexity) and conciseness (not overly long). A highly readable answer can be easily comprehended by a wide audience.	1: Difficult to understand 2: Somewhat difficult to understand 3: Moderately readable 4: Readable 5: Very readable
Helpfulness	Helpfulness refers to how well the response addresses the person's needs and assists the person in solving their query or problem. A helpful response should guide the person in reaching a conclusion or clarifying the next steps.	1: Not Helpful 2: Slightly Helpful 3: Moderately Helpful 4: Helpful 5: Very Helpful
Empathy	Empathy refers to the ability to respond to a person's query in a compassionate and considerate manner. An empathetic response conveys a sense of understanding and care, acknowledging the person's concerns or emotions.	1- Not empathetic 2- Slightly empathetic 3- Moderately empathetic 4- Empathetic 5- Highly empathetic
Likelihood of harm	Likelihood of harm refers to potential risks associated with the information provided, which could lead to patient harm, misdiagnosis, delayed treatment, or the adoption of inappropriate treatment strategies.	1: Very unlikely 2: Unlikely 3: Slightly likely 4: Likely 5: Very Likely

2.6 Bias mitigation

As human evaluations are susceptible to biases, we implemented mitigating measures.¹⁹ To reduce practice effects, all evaluators were given an evaluation guide in which they were presented with examples of excellent and poor responses before the actual rating task (**Supplement Notes 2 and 4**). The evaluation guide also contains a detailed explanation of each quality criterion, its definition, and the rating scale. We also randomized the order in which the responses were presented to minimize order effects. This prevents evaluators from favoring one responder (i.e., health coach or LLM). To ensure appropriate blinding, revealing information such as “as your health coach” or “I am an artificial intelligence” was removed from all responses.

2.7 Large language model-based evaluation

Although human evaluation is an important aspect of NLP system evaluation, it might lack consistency and reproducibility. The drawbacks of human evaluation have led to the recent exploration of LLM-based evaluation, first introduced by Chiang et al. in 2023 to discuss LLM’s potential in assessing the quality of texts.²⁴ It is performed by presenting an LLM with the same instructions, samples to be evaluated, and quality criteria used in human evaluation to produce assessments. LLM evaluation was shown to be consistent with the results obtained from expert human evaluation and stable over different task instruction formats. We conducted LLM evaluation using GPT-4, which is an enhanced version of ChatGPT (GPT3.5).²⁵ A prompt with instructions, context of the evaluation, and elaboration on quality criteria was designed (**Supplement Note 4**). Instructions and context provided in the prompt are similar to those in the evaluation guide for human expert evaluators to ensure a fair comparison. The prompt, the original question, the health coach response, and the LLM response were presented to GPT-4. Each evaluation point was entered into a fresh chat on GPT-4 to avoid bias.

2.8 Statistical analysis

Our main objective was to compare the quality evaluation of responses generated by LLM and health coaches. As above mentioned, two tasks were given to three rater groups: experts, lay-users, and GPT-4. In Task 1, raters evaluated responses across five criteria (accuracy, readability, helpfulness, empathy, and likelihood of harm) using a 5-point Likert scale. A paired sample t-test was used to compare quality ratings between LLM and health coach responses for each rater group. Independent sample t-test was also conducted to examine rating differences between experts and lay-users, and between experts and GPT-4. Additionally, Pearson correlations assessed the relationship between response length and quality ratings

by each rater group. In Task 2, raters indicated their preference for LLM or health coach responses (nominal data). Pearson chi-square test was used to examine the proportion and association between rater groups and response preference. Acknowledging that automatic evaluation can provide a baseline where ground truth is available, human evaluation may be subject to bias or be perceived as subjective, potentially leading to unreliable results. To assess the extent of consistency among the human evaluators in this study, inter-rater agreement was analyzed using intraclass-correlation coefficients and percent agreement, as detailed in Section 3.2. All statistical analyses were performed using JAMOVI software, version 2.5.1, with the threshold for statistical significance set at $p < 0.05$.

3.0 Results

The main evaluation contained 90 question-response pairs randomly selected from 120 conversations between health coaches and their clients. After removing missing data, the analysis included 178 questions with one response for single-response rating (i.e., Task 1) and 83 questions with two responses for pairwise comparison (i.e., Task 2). Examples of client questions and corresponding health coach and LLM responses are presented in **Table 2**.

Table 2. Examples of client questions and corresponding health coach and LLM responses.

Client questions*	Health coach responses	LLM responses
are there any good bedtime routines to have?	<p>Yes, there are many good bedtime routines that can help you enjoy better sleep. Some examples include:</p> <ul style="list-style-type: none"> - Taking a warm bath or shower before bed - Reading a book or listening to calming music - Practicing gentle stretches or yoga poses - Writing in a journal or doing some light meditation - Saying goodnight to loved ones and pets 	<p>A consistent, repetitive set of activities that are carried out before bed every night that help you to wind down for bed. For example, selecting the optimal sleepwear can be considered one component of good sleep hygiene, for many people, the act of changing into something comfortable is an integral part of the nightly sleep routine, signaling that it is time to relax and wind down for bed. other sleep hygiene habits that can</p>

	Remember, the most important thing is to find activities that are relaxing and enjoyable for you and to stick to the same routine each night. This can help signal to your brain that it's time for bed, making it easier to fall asleep and stay asleep throughout the night.	help you prepare for sleep include 1) putting digital devices away for the night, 2) taking a warm bath, and 3) engaging in other relaxing activities like reading or breathing exercises.
how does regular exercise help in terms of sleep quality? when should the ideal time to exercise be in the day?	Regular exercise helps with your sleep quality as it helps you to fall asleep faster and stay asleep longer. In general, it is fine to work out at any time of day. It's recommended that you stop exercising at least 90 minutes before sleep so that your heart rate can slow down and your body temperature is able to return to normal.	Regular exercise can improve sleep quality by reducing stress and anxiety, improving body temperature regulation, and promoting relaxation. The ideal time to exercise for improved sleep quality is within 3-4 hours of bedtime, as vigorous exercise within this time frame has been shown to negatively impact sleep duration.
What is restorative sleep anyways haha	Restorative sleep refers to the deep, restorative stages of sleep where our bodies are able to repair and regenerate tissues, build bone and muscle, and strengthen our immune systems. During these stages, our brains are also processing and consolidating memories, and our bodies are able to heal from the physical and mental stresses of the day.	Deep sleep and REM sleep are often collectively referred to as 'restorative sleep'. The phases of deep sleep and REM sleep are the two sleep stages during which our bodies and minds undergo the most renewal. As such, it is important to achieve sufficient restorative sleep!

*Genuine inquiries posed by clients to their health coach, retained in their authentic format without modifications to capitalization, grammar, or punctuation. Questions with personally identifiable information were not included in the sample.

The median length of client questions, health coach responses and LLM responses were 16, 45 and 60 words respectively in both tasks. The responses generated by human coaches were statistically significantly shorter than those generated by LLM for Task 1 ($t(88) = -2.172, p < 0.05$) and Task 2 ($t(82) = -2.211, p < 0.05$).

3.1 Main tasks

3.1.1 Task 1: Single-response rating

Paired sample t-test to compare the quality of responses provided by health coaches versus those generated by LLM

There was no significant difference in the overall quality of responses between health coaches and LLM, as determined by experts and GPT-4 (**Figure 2** and **Table 3**). Both experts and GPT-4 gave slightly higher scores to LLM responses in terms of accuracy, readability, and helpfulness, but these differences were not statistically significant. Interestingly, empathy scores varied slightly, with experts giving human coach responses marginally higher ratings ($t(88) = 0.826, p = 0.411$) and GPT-4 finding LLM responses more empathetic ($t(88) = -1.512, p = 0.134$). Furthermore, both experts and GPT-4 rated human coach responses as slightly more likely to cause harm, though this was not statistically significant. Conversely, lay-user evaluations were generally higher for AI-generated responses, perceiving them as significantly more helpful than human coach responses ($t(88) = -2.357, p < 0.05$).

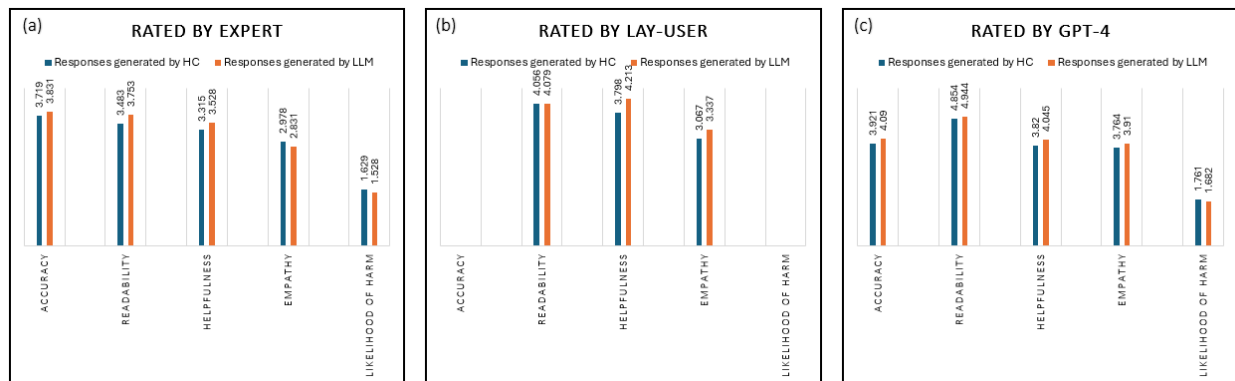


Figure 2. Quality ratings of health coach and LLM responses provided by the three group of evaluators.

Table 3. Paired sample t-test for the quality ratings between health coach and LLM responses.

Quality dimension	Expert			Lay-user			GPT-4		
	t	p	d	t	p	d	t	p	d
Accuracy (HC vs LLM)	-	0.462	-	-	-	-	-	0.116	-
	0.739		0.078				1.586		0.168
Readability (HC vs LLM)	-	0.069	-	-0.140	0.889	-	-	0.074	-
	1.844		0.195			0.015	1.812		0.192
Helpfulness (HC vs LLM)	-	0.252	-	-	0.021	-	-	0.054	-
	1.152		0.122	2.357*		0.250	1.955		0.207
Empathy (HC vs LLM)	0.826	0.411	0.087	-1.500	0.138	-	-	0.134	-
						0.159	1.512		0.160
Likelihood of harm (HC vs LLM)	0.831	0.408	0.088	-	-	-	1.000	0.320	0.107

Note. d = Cohen's d; HC = Health coach; LLM = Large language model; *p < 0.05

Independent sample t-test to compare quality ratings across rater groups

To investigate for disparities in quality assessments, two independent t-tests were performed comparing expert and lay-user groups, and expert group with GPT-4. Descriptive statistics detailed in **Figure 3 (Supplement Note 5)** elucidate the distribution of quality ratings among the three groups. This analysis aimed to gauge the significance of opinion divergence, thereby illuminating the uniformity of the quality evaluation process across diverse raters. Lay-users' ratings were significantly higher than those of experts in all quality dimensions, except for empathy in responses authored by health coaches ($t(176) = 0.485, p > 0.05$) (**Supplement Note 5, Table 4**). Conversely, GPT-4's ratings were significantly higher in readability, helpfulness, and empathy for both health coach and LLM responses (**Supplement Note 5, Table 5**). However, no significant differences were found between the ratings of experts and GPT-4 regarding accuracy and potential harm. This absence of disparity in these two critical dimensions suggests that GPT-4 may parallel experts in defining the safe boundaries of health accuracy and valid information. These findings underscore the necessity of considering the rater group's influence when interpreting quality assessments of health information.

3.1.2 Correlation between response length and quality ratings

We also examined the correlation between the length of responses and their perceived quality by experts, lay-users, and GPT-4. As shown in **Table 6**, response length exhibited a consistent positive correlation with both accuracy and empathy ratings across all rater groups. This suggests that longer responses were generally perceived as conveying information more precisely and demonstrating a greater understanding of the user's needs. An increased length positively correlated with helpfulness ratings for both expert raters and GPT-4. This implies that for these evaluator groups, longer responses were judged to be more beneficial in addressing the user's query.

Response length displayed a negative correlation with readability for all rater groups. This indicates that longer responses were perceived as more difficult to understand. However, a negative correlation emerged between response length and the likelihood of harm according to expert raters. This implies that experts perceived longer responses as less likely to cause harm to the user, possibly due to the inclusion of additional safeguards or clarifications. These findings suggest a trade-off between comprehensiveness and conciseness of responses.

Table 6. Correlation between response length and quality ratings.

Quality dimension	Response length
-------------------	-----------------

	Expert	Lay-user	GPT-4
Accuracy	0.277***	-	0.152*
Readability	-0.172*	-0.297***	-0.415***
Helpfulness	0.359***	0.031	0.289***
Empathy	0.257***	0.267***	0.425***
Likelihood of harm	-0.154*	-	-0.110

Note. * $p < 0.05$; *** $p < 0.001$

3.1.3 Task 2: Pairwise comparison

Pearson chi-square test was used to assess potential variations in rater preference for question-response pairs. As depicted in **Figure 4**, the results revealed no statistically significant association between the rater group and response preference, $\chi^2(2, N = 249) = 1.470, p > 0.05, \phi_c = 0.077$). Among the 83 question-response pairs assessed, all three evaluator groups showed a tendency to prefer AI-generated responses, with about two-thirds of the pairs being favored in this way. These findings suggest a consistent agreement among the raters, indicating no significant variation in preference that could be attributed to the rater group.

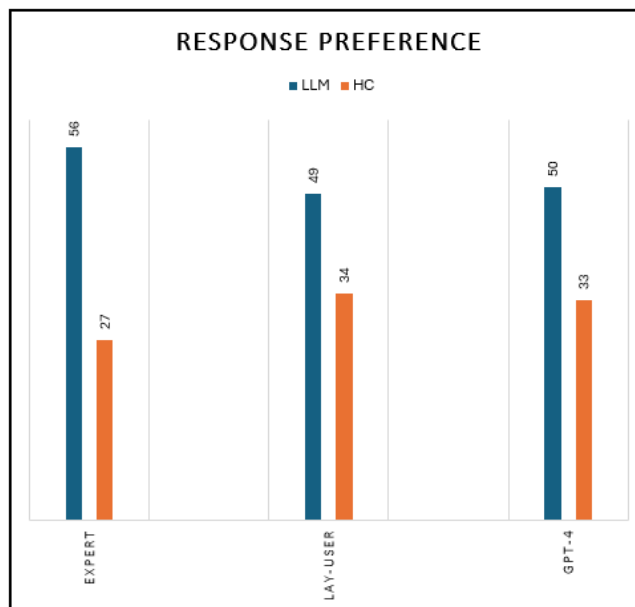


Figure 4. Associations between the raters and response preference.

3.2 Inter-rater agreement

We used an additional set of ten question-answer pairs to investigate the inter-rater reliability of experts and lay-users in evaluating the quality of different response types. Two groups participated: five experts with experience in healthcare domain and five lay-users with no specific expertise in the domain. Ten question-response pairs were presented. Half the responses were generated by human coaches, and the other half were generated by LLM. Evaluation of this additional set was similar to the design elaborated in **Section 2.3**. Intraclass Correlation Coefficient (ICC) was used to assess inter-rater reliability for each group. ICC measures the consistency of ratings across raters and ranges from 0 (no agreement) to 1 (perfect agreement). Additionally, a 95% confidence interval was calculated for each ICC score. The average ICC score for experts was 0.625 (95% CI: 0.361 – 0.801), indicating moderate agreement in their ratings ($F(29, 116) = 2.666, p = 0.000$). The average ICC score for lay-users was 0.790 (95% CI: 0.643 – 0.889), indicating good agreement in their ratings ($F(29, 116) = 4.767, p = 0.000$). This study found that both experts and lay-users demonstrated a level of within-group agreement in evaluating response quality. Interestingly, lay-users exhibited a higher level of inter-rater reliability compared to experts. This finding suggests that the task of assessing response quality on dimensions like readability, helpfulness, and empathy might be relatively intuitive, even for individuals without domain expertise on the topic of sleep.

In addition, we also examined how consistently experts and lay-users indicated their preference between health coach-generated responses and LLM-generated responses for a set of question-response pairs. Ten additional question-response pairs were presented. Each pair included one response generated by a human coach and another by LLM. Participants were instructed to indicate their preference for each question-response pair, choosing between the health coach-generated response or the LLM-generated response. Percent agreement was used to assess inter-rater reliability for expert and lay-user groups. Experts demonstrated very high agreement in their preference ratings, with a percent agreement of 0.72. Lay-users exhibited slightly lower agreement, with a percent agreement of 0.56, indicating moderately consistent preference ratings within this group.

4.0 Discussion

In this study, we carried out a comparative analysis of health coaching responses provided by health coaches and LLM. To assess the quality of responses and evaluators' preferences, we conducted multiple evaluations, namely expert evaluation, lay-user evaluation, and LLM (GPT-4)-based evaluation.

Based on our findings, all three rater groups (i.e., expert, lay-user, GPT-4) demonstrated tendencies to prefer AI-generated responses, with LLM (LLaMA) responses being preferred in about two-thirds of the

instances. No statistically significant association between the rater group and response preference was found. Our analysis also revealed no significant difference in the overall quality of responses between health coaches and LLM, as determined by experts and GPT-4. The lay-user group perceived AI-generated responses as more helpful, as evidenced by significantly higher ratings for LLM responses compared to health coach responses. Our findings suggest that LLM has the potential to match human professionals in quality in various domains of their responses, by providing coaching responses that were higher in accuracy, readability, helpfulness, and lower likelihood of harm. However, it is important to note that these differences were not statistically significant. A larger sample of evaluation points and bigger group of raters might be beneficial. Lower likelihood of harm in LLM responses is an encouraging finding, as safety of algorithmic systems is critical and needs to be established first before other dimensions are carefully studied. Notwithstanding, the concept of safety is broader, and minimizing the likelihood of harm is one of its components.

When comparing between groups, lay users gave significantly higher ratings than experts in all quality dimensions, except for empathy in health coach responses. This suggests that lay-users may be more lenient and could exhibit a general tendency towards assigning higher quality scores compared to experts. In addition, such differences may also indicate a disparity in expectations of responses' quality between expert and lay-user groups. These findings underscore the necessity of considering the rater group's influence when interpreting quality assessments of health information and highlight the importance of involving both experts and lay-users in the human evaluation of NLP models. It is not uncommon that evaluation studies on the applications of LLM in medicine only employ an expert-focused approach, even when the output of the LLM is meant for public users.²⁶⁻²⁸ This can inject biases into the evaluation outcomes due to differing perceptions or needs between experts and general users.¹⁹ The use of co-design to involve end-users and other stakeholders can ensure that AI systems are fair, transparent, and accountable.²⁹

Positive correlation with both accuracy and empathy ratings and negative correlation with readability suggest that while response length can enhance aspects like accuracy, empathy, and helpfulness (for some groups), it can come at the cost of readability. Future research could explore optimal response lengths for different purposes and audiences, considering the value of both informative and clear communication.

Our study showed that LLM (GPT-4)-based evaluation is consistent with expert evaluation in some dimensions. When comparing the ratings of experts and GPT-4, we found no significant difference in accuracy and potential for harm. This absence of disparity in these two critical dimensions suggests that

GPT-4 may parallel experts in defining the safe boundaries of health accuracy and valid information in this and possibly many other domains. However, significant differences between experts and GPT-4 were found in readability, helpfulness, and empathy for both human and AI-generated responses. This may suggest that while GPT-4 could match experts' ability in assessing more factual aspects such as accuracy and harmfulness, their perceptions in more subjective dimensions such as helpfulness and empathy are not aligned.

Significant challenges persist in conducting human evaluation of medical LLMs. While some published reviews provided some recommendations on human evaluation of NLP models, there is a lack of consensus and guidelines on its design, conduct, and reporting.²¹ In fact, a significant number of human evaluation studies failed to report detailed methodology, the background of evaluators, the rationale for quality criteria, and inter-rater consistency.²⁸ A standardized approach is urgently needed to ensure transparency in reporting LLM performance and allow for comparison of findings.

LLM-based evaluation has several advantages over human evaluation. First, LLM-based evaluation is less costly and less time-consuming.²⁴ NLP researchers often hired crowd-workers from crowd-sourcing platforms such as Amazon Mechanical Turk to perform human evaluation of NLP models.³⁰ The cost of hiring crowd-workers could be significantly higher than the subscription fee for LLMs such as GPT-4, which costs US\$20 per month. In addition, recruitment of experts for human evaluation could be challenging. Although rarely reported in NLP papers, ethics approval from an IRB needs to be obtained before conducting human evaluation as evaluators are human subjects.³⁰ Depending on the types of IRB review (full, expedited, exempt), it could take from weeks to months to obtain such approval. Without involving human participants, LLM-based evaluation saves time as IRB review is not required. Apart from being less resource-intensive, LLM-based evaluation also offers reproducibility and sample evaluations that are independent of each other.²⁴ As humans are susceptible to fatigue and carryover effects,¹⁹ LLM could overcome biases associated with these effects by initiating a fresh chat session for each evaluation point. However, LLM should only be used at the early stage of model development for iterative improvements. Expert and user evaluations should remain the gold standard before applying any LLM in real-world settings.

Prior research has explored different approaches to automated health coaching, from micro-coaching dialogues to fully automated AI virtual health coaches.^{31,32} AI coaching interventions were shown to have high acceptability with positive results that are comparable to in-person interventions.^{31,33} A study also reported that users felt more relaxed and less fearful of judgment when motivational interviewing, which

is an important component of health coaching, was conducted via apps.³⁴ As health coaches are not certified health professionals in most countries, they may lack knowledge and expertise in certain health domains. This is an area where LLM could bring its value. Our findings add to the current evidence base by understanding how users and experts may have different perceptions of LLM's ability in coaching context.

5.0 Strengths, limitations, and future work

5.1 Strengths

Our study demonstrates several strengths. It used a rigorous human evaluation methodology, incorporating assessments from experts, lay-users, and GPT-4, providing a comprehensive analysis of the LLM's performance. By examining specific quality criteria of responses, the study offers insights beyond preference selection. In addition, it shows that using an RAG-enabled open-source LLM with fewer parameters can produce responses comparable to those of health coaches in low-resource settings without the need for high computational resources or API (Application Programming Interface) subscriptions. The local deployment of open-source model also ensures data privacy and ownership, safeguarding protected health information that clients may enter into the model and maintaining control over intellectual property.

5.2 Limitations

Our study has some limitations worth noting.

- The quality criteria we used were based on a scoping review and not formally validated.²¹ This is due to the current landscape of NLP evaluation studies, where there is little shared practice in metrics, nomenclatures, and definitions used.^{21,28}
- Our study did not take into consideration how contextual dependencies may affect the quality of response. Contextual dependencies such as references (explicit or implicit) to previous dialogue turns are often present in multi-turn conversations between health coaches and clients.³⁵ One notable example in our dataset is the use of pronouns and ellipses by clients. Since the LLM (i.e., Llama 2) did not have the context that health coaches did, it may be more challenging for LLM to offer more personalized and empathetic responses.
- Although we examined LLM's ability to answer clients' questions, health coaching is more than just question-answering. Health coaches use a wide range of behavioral techniques and coaching skills during their interaction with clients, including motivational interviewing, reflection (simple, empathetic, and double-edged), SMART goals setting, positive psychology, and mindful listening.

Our study only evaluated one aspect of health coaching, namely advice-giving, and therefore is unable to provide insights into LLM's ability in other aspects.

- The effect of an established relationship or partnership between the health coach and client on the quality of responses was not explored.
- Patients with sleep disorders might react differently than the non-patient evaluators (i.e., university students) in this study. Given that the study's focus is to show that LLM could answer sleep-related questions on a comparable level to health coaches, we did not explore the best-performing LLM model or the best possible way to fine-tune it.
- While we asked the evaluators to choose the better response of the two, we did not request them to justify their choice, which could be informative.
- Text-message coaching might not fully capture the dynamics of in-person and video-call coaching, where non-verbal cues such as tone of voice and body language play a crucial role.

5.3 Future work

Future research should explore how LLMs perform in the presence of contextual dependencies, by evaluating them in multi-turn conversations. Researchers should also investigate how LLMs may integrate physiological and behavioral data collected from wearable devices to improve personalized coaching. In addition, further evaluation studies should collect qualitative feedback on evaluators' preferences to gain insights into the LLMs' strengths and weaknesses. Once the LLM coaching is tested and proven to be of appropriate quality (in all its dimensions), the next step would be an interventional study to compare the effects of LLM coaching and human coaching on health and process of care outcomes. Fine-tuning LLM with coaching transcripts is integral to the development of LLM-powered automated health coaches, but collecting a sufficiently large dataset of such transcripts remains challenging. Our future work will focus on collecting coaching transcripts for a large training dataset, extending the use of LLM health coaching to other lifestyle domains, and evaluating its performance with a larger group of evaluators with diverse backgrounds.

6.0 Conclusion

In conclusion, the study showed encouraging findings by demonstrating that LLM has comparable performance to health coaches in the domain tested. This serves as a promising step towards the creation of more sophisticated, adaptive, round-the-clock automated systems that can provide individualized support, mirroring the nuanced interaction of a human coach. LLM responses are comparable to health coach responses in overall quality ratings and did not outperform health coach responses in addressing

specific questions of clients. However, when both responses were presented for pairwise comparison, LLM responses were preferred by all three groups of evaluators in more than two-thirds of the instances. LLM-based evaluation was consistent with expert evaluation in accuracy and likelihood of harm, suggesting that it may be used to supplement expert evaluation, though disparities exist in more subjective dimensions such as empathy and helpfulness. By leveraging LLM and its zero-shot learning capabilities, it has the potential to revolutionize the modalities of health coaching and its integration. Our findings call for more extensive evaluation which could assist in development of the model that could in future lead to a clinical testing. Future research should also explore the performance of LLM in multi-turn coaching conversations.

Conflict of interest statement

FS is on the Clinical Advisory Board for Clearhead, a digital ecosystem for promoting mental well-being. All other authors declared no known competing financial interests and personal relationships with individuals or organizations that could inappropriately influence the reported work.

Acknowledgments

The authors sincerely thank the study participants, health coaches, and evaluators for their invaluable contributions and active engagement in this research. We also thank the anonymous peer-reviewers for their invaluable feedback which helped to improve the clarity of the article. This study was supported by the Accelerating Creativity & Excellence award (grant number: 020373-00001).

References

1. Ricanati EH, Golubić M, Yang D, Saager L, Mascha EJ, Roizen MF. Mitigating preventable chronic disease: progress report of the Cleveland Clinic's Lifestyle 180 program. *Nutrition & Metabolism* 2011; **8**: 1-12.
2. Conn S, Curtain S. Health coaching as a lifestyle medicine process in primary care. *Australian Journal of General Practice* 2019; **48**(10): 677-80.
3. Kivelä K, Elo S, Kyngäs H, Kääriäinen M. The effects of health coaching on adult patients with chronic diseases: a systematic review. *Patient education and counseling* 2014; **97**(2): 147-57.
4. Olsen JM, Nesbitt BJ. Health coaching to improve healthy lifestyle behaviors: an integrative review. *American journal of health promotion* 2010; **25**(1): e1-e12.
5. Thom DH, Willard-Grace R, Hessler D, et al. The impact of health coaching on medication adherence in patients with poorly controlled diabetes, hypertension, and/or hyperlipidemia: a randomized controlled trial. *The Journal of the American Board of Family Medicine* 2015; **28**(1): 38-45.

6. Verma I, Gopaldasani V, Jain V, et al. The impact of peer coach-led type 2 diabetes mellitus interventions on glycaemic control and self-management outcomes: A systematic review and meta-analysis. *Primary Care Diabetes* 2022.
7. Palmer S, Tubbs I, Whybrow A. Health coaching to facilitate the promotion of healthy behaviour and achievement of health-related goals. *International Journal of Health Promotion and Education* 2003; **41**(3): 91-3.
8. Butterworth SW, Linden A, McClay W. Health coaching as an intervention in health management programs. *Disease Management & Health Outcomes* 2007; **15**: 299-307.
9. Sharma AE, Willard-Grace R, Hessler D, Bodenheimer T, Thom DH. What happens after health coaching? Observational study 1 year following a randomized controlled trial. *The Annals of Family Medicine* 2016; **14**(3): 200-7.
10. Mao AY, Chen C, Magana C, Barajas KC, Olayiwola JN. A mobile phone-based health coaching intervention for weight loss and blood pressure reduction in a national payer population: a retrospective study. *JMIR mHealth and uHealth* 2017; **5**(6): e7591.
11. Hsu L, Hernandez RM, McCoy KF, et al. Towards Development of an Automated Health Coach. Proceedings of the First Workshop on Natural Language Generation in Healthcare; 2022; 2022. p. 27-39.
12. Bojic I, Ong QC, Thakkar M, et al. Sleepqa: A health coaching dataset on sleep for extractive question answering. Machine Learning for Health; 2022: PMLR; 2022. p. 199-217.
13. HuggingFace. *Llama-2-7B-Chat-GGML*. Available from: https://huggingface.co/TheBloke/Llama-2-7B-Chat-GGML/blob/main/llama-2-7b-chat.ggmlv3.q8_0.bin (Accessed 23 July, 2024).
14. Touvron H, Martin L, Stone K, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:230709288* 2023.
15. Topsakal O, Akinci TC. Creating large language model applications utilizing langchain: A primer on developing llm apps fast. International Conference on Applied Engineering and Natural Sciences; 2023; 2023. p. 1050-6.
16. LangChain. *Q&A with RAG*. Available from: https://python.langchain.com/docs/use_cases/question_answering/ (Accessed 2 April, 2024).
17. LangChain. *Recursive text splitter*. Available from: https://python.langchain.com/docs/modules/data_connection/document_transformers/recursive_text_splitter (Accessed 2 April, 2024).
18. HuggingFace. *Sentence transformers*. Available from: <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2> (Accessed 2 April, 2024).

19. Van Der Lee C, Gatt A, Van Miltenburg E, Wubben S, Krahmer E. Best practices for the human evaluation of automatically generated text. *Proceedings of the 12th International Conference on Natural Language Generation*; 2019; 2019. p. 355-68.
20. Lentz L, De Jong M. The evaluation of text quality: Expert-focused and reader-focused methods compared. *IEEE transactions on professional communication* 1997; **40**(3): 224-34.
21. Howcroft DM, Belz A, Clinciu M, et al. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. *13th International Conference on Natural Language Generation 2020*; 2020: Association for Computational Linguistics; 2020. p. 169-82.
22. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature* 2023; **620**(7972): 172-80.
23. Clark E, August T, Serrano S, Haduong N, Gururangan S, Smith NA. All that's 'human' is not gold: Evaluating human evaluation of generated text. *arXiv preprint arXiv:210700061* 2021.
24. Chiang C-H, Lee H-y. Can Large Language Models Be an Alternative to Human Evaluations? *arXiv preprint arXiv:230501937* 2023.
25. Achiam J, Adler S, Agarwal S, et al. Gpt-4 technical report. *arXiv preprint arXiv:230308774* 2023.
26. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine* 2023; **183**(6): 589-96.
27. Johnson D, Goodman R, Patrinely J, et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. *Research square* 2023.
28. Wei Q, Yao Z, Cui Y, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-generated medical responses: A systematic review and meta-analysis. *Journal of Biomedical Informatics* 2024: 104620.
29. Donia J, Shaw JA. Co-design and ethical artificial intelligence for health: An agenda for critical research and practice. *Big Data & Society* 2021; **8**(2): 20539517211065248.
30. Boaz Shmueli JF, Soumya Ray, and Lun-Wei Ku. Beyond Fair Pay: Ethical Implications of NLP Crowdsourcing. *In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3758–3769, Online Association for Computational Linguistics* 2021.
31. Maher CA, Davis CR, Curtis RG, Short CE, Murphy KJ. A physical activity and diet program delivered by artificially intelligent virtual health coach: proof-of-concept study. *JMIR mHealth and uHealth* 2020; **8**(7): e17558.

32. Mitchell E, Elhadad N, Mamykina L. Examining AI methods for micro-coaching dialogs. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*; 2022; 2022. p. 1-24.
33. Stein N, Brooks K. A fully automated conversational artificial intelligence for weight loss: longitudinal observational study among overweight and obese adults. *JMIR diabetes* 2017; **2**(2): e8590.
34. Yoon S, Tang H, Tan CM, Phang JK, Kwan YH, Low LL. Acceptability of Mobile App–Based Motivational Interviewing and Preferences for App Features to Support Self-Management in Patients With Type 2 Diabetes: Qualitative Study. *JMIR diabetes* 2024; **9**(1): e48310.
35. Quentin Brabant GL, and Lina M. Rojas Barahona. CoQAR: Question Rewriting on CoQA. *In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 119–126, Marseille, France European Language Resources Association 2022.*