Target: PLoS Digital Health 1–12

Infusing behavior science into large language models for activity coaching

Madhurima Vardhan*
Narayan Hegde* HEGDE@GOOGLE.COM
Deepak Nathani
Google Research, Bangalore, India
Emily Rosenzweig
Verily Life Sciences, San Francisco, USA
Alan Karthikesalingam
Martin Seneviratne
Google Health, London, UK
* *equal first*

Abstract

Large language models (LLMs) have shown promise for task-oriented dialogue across a range of domains. The use of LLMs in health and fitness coaching is under-explored. Behavior science frameworks such as COM-B, which conceptualizes behavior change in terms of capability (C), Opportunity (O) and Motivation (M), can be used to architect coaching interventions in a way that promotes sustained change. Here we aim to incorporate behavior science principles into an LLM using two knowledge infusion techniques: coach message priming (where exemplar coach responses are provided as context to the LLM), and dialogue re-ranking (where the COM-B category of the LLM output is matched to the inferred user need). Simulated conversations were conducted between the primed or unprimed LLM and a member of the research team, and then evaluated by 8 human raters. Ratings for the primed conversations were significantly higher in terms of empathy and actionability. The same raters also compared a single response generated by the unprimed, primed and re-ranked models, finding a significant uplift in actionability from the re-ranking technique. This is a proof of concept of how behavior science frameworks can be infused into automated conversational agents for a more principled coaching experience.

Institutional Review Board (IRB) The study does not involve human subjects beyond the volunteer annotators. IRB approval was not sought for this research.

1. Introduction

It is estimated that 81% of adolescents and 27% of adults do not achieve the levels of physical activity recommended by the World Health Organization (WHO) (1). A sedentary lifestyle is associated with long term adverse health outcomes, ranging from cardiovascular disease and diabetes to mental health problems and cognitive decline (2). A 2022 report found that progress toward these goals has been slower than expected and highlighted digital health tools as a particular opportunity area (3).

Numerous smartphone nudging tools have been designed to promote physical activity (4; 5). These interventions are low-cost and highly scalable relative to human fitness coaches, with promising early evidence (6; 7; 8; 9). One randomized controlled trial of a digital walking coach found short-term improvements in physical activity (10). However, in an era of notification overload, there is also a risk of desensitization and alert fatigue if the nudge strategy is not well designed.

Automated conversational agents offer an opportunity to create interactive dialogue, with widespread applications in e-commerce, home automation and healthcare (11; 12). Health and Fitness coaching is emerging as a promising use case for these conversational agents (13; 14; 15; 16). However, most traditional systems are limited in their degree of personalization and persuasiveness because they depend on rule-based

nudge engines with static message content rather than adaptive conversational agents that can mimic realistic dialogue from a human coach (17).

Large language models (LLMs), such as GPT-3 (18), PaLM (19), Gopher (20) and LaMDA (21), excel in natural language generation with greater expressivity and versatility compared to rule-based chatbots. To date, use of LLMs in the health and fitness space has been limited, however interest is growing rapidly following the release of LLMs tailored to biomedical tasks (22). A major challenge in using LLMs in health care is how to ensure the model is personalized and adaptive while still remaining consistent with evidence-based practice and within safety guardrails (23). Activity coaching relies on complex interpersonal dynamics where the coach builds rapport with the trainee, provides motivation, helps to overcome pre-existing patterns of behavior, etc.- which are not explicitly optimized in LLMs (24). Knowledge infusion refers to the integration of established knowledge or practice into a model. In principle this is often achieved via finetuning on a task-specific dataset (25). The disadvantage of finetuning in the coaching domain is that it requires coaching transcripts, which are difficult to obtain. Finetuning has also been shown to diminish the few-shot performance of a pretrained LLM with in-context prompts - i.e. over-specialization of the model (26). Knowledge infusion is an active area of research and many other methods exist including customizing training objectives (27), reinforcement learning with human feedback (28; 29), in-context learning via prompt engineering or priming (30; 31) and many associated prompt design variants (32; 33; 34; 35). There have also been numerous strategies to ensemble knowledge infusion techniques, including post-hoc re-ranking or summarization of model outputs to further align the model with the task of interest (36; 37). Customizing knowledge infusion strategies for the health care domain remains an area of active research. Here we propose two simple in-context learning methods to infuse behavior science principles into LLMs without the requirement for finetuning or reinforcement learning.

Coaching in the context of physical activity ranges from delivering tailored products that serve elite athletes, to creating motivational tools that support inactive users to become fitter through a progressive and personalised programs. Our LLM is designed to target latter use case to help users lead more active lifestyle using behavioral nudges and resolving barriers through conversations.

Behavioral science offers theoretical frameworks to help understand the factors influencing human behavior and design effective behavior change interventions for a given context. COM-B is a well-known framework which conceptualizes behavior change along three axes: Capability (the psychological and physical skills to act); Opportunity (the physical and social conditions to act); and Motivation (the reflective and automatic mental processes that drive action) (38). Behavioral science can be useful to guide the design of automated nudging systems for habit formation (39).

We extend the PACE (16) work on designing automated physical activity coaching engine based on an analogous behavior science framework called Fogg's Behavior Model. A rule-based automated nudging agent based on this model had comparable outcomes to human coaches in terms of user step count and engagement. In this study, we extend findings of the PACE study by connecting the strengths of a behavioral science rule-based model with the conversational versatility of an LLM. The goal is to address the broader question of how behavior science principles might guide or constrain conversational LLMs. Specifically, we make use of priming and dialogue re-ranking. These are both lightweight techniques that do not require additional model retraining or finetuning. Overall, the key contributions contributions of this study are as follows:

1. Defining evaluation metrics for LLM conversations in the activity coaching domain
2. Introducing two different approaches to behavioral science knowledge infusion: coach phrase priming and dialogue re-ranking
3. Evaluating the benefit of knowledge infusion relative to an unprimed LLM using quantitative and qualitative approaches

2. Methods

The following sections outline the datasets, language modeling techniques and evaluation methods used.

2.1. Data

The previous PACE study dataset was re-purposed for this analysis (16). Specifically, this dataset was used to construct the example coaching phrases used in the behavior science priming, create training data for finetuning BERT user and coach statement classifiers and to select the user queries (initial user responses) in simulated conversations for evaluation. This dataset consists of dialogue transcripts between fitness coaches and subjects, generated from real coaching interactions across various activity habit formation related issues. In this Wizard-of-Oz study design, consented subjects were randomized to coaches or coaches using a FBM assistant that suggested example responses based on behavior science using a rule-based engine. The dataset included 520+ conversations from 33 participants over 21 days. A total of 6 independent annotators labeled these conversations as one of Motivation, Capability and Opportunity. Both user and coach statements where separately annotated with presence or absence of each of these three themes. Data collection and annotation protocol is described in detail in (16).

2.2. Language models

The Language Models for Dialog Applications (LaMDA) pretrained LLM was used as the primary architecture (21), with no further finetuning. LaMDA is a decoder-only transformer architecture with 64 layers, used here in its 137 billion parameter configuration. We used the following LaMDA hyperparameters: temperature 0.9; maximum token length 1024, top k (controls sampling diversity) 40. LaMDA has an option to provide context alongside the LLM prompt - this was how the coach phrase priming was conducted. LaMDA also provides top-k outputs, which were used in the re-ranking (see below).

2.3. Coach phrase priming

Coach phrase priming was performed by inputting 30 example coach nudges as context to the LLM prior to the prompt. The 30 nudges were selected from the data in the PACE study - specifically the 10 most common coach responses in each of the three behavior science categories of interest: C/O/M. Details regarding coach phrase selection and priming method are described in section 1 of supplementary paper. For example, the Capability category included activity planning and barrier conversations; and Opportunity included social engagement conversations and activity planning; and Motivation included congratulations and positive affirmation; [ref]. The order of the 30 nudges was randomized. The priming prompts are shown in Table 1.

2.4. Simulated dialogue

The following LLM configurations were compared via simulated conversations with a single member of the research team:

1. Unprimed (trigger prompt only)

2. Coach-primed (30 example nudges provided as LLM context)

All conversations begin with the trigger prompt: *Hey John, It's time for your morning walk.* The subsequent user responses were sampled from a set of 9 user statements, with 3 each designed to evoke a low Motivation, low Capability and low Opportunity (user queries are included in the Supplementary Materials table 2). An example user statement with low opportunity was: *I am super busy with work today. I have chores to do in the morning and work meetings after that.*.

This culminated in a total of 18 transcripts: 9 each for the unprimed and primed LLMs. The conversations were continued with dialogue between the LLM and the human interlocutor (researcher). The conversations were terminated at a natural breakpoint at the discretion of the researcher. Any follow up questions to the LLM response were added appropriately to continue the conversation on the original topic until a logical end was reached. Additional example transcripts are contained in the Supplementary Materials.

Table 1: LLM prompts used in coach phrase priming.

Behavior Science Priming
The following is a conversation with an AI Health Coach. The coach tries to motivate the users when the user lacks motivation, can resolve barriers. Here are some examples of how a coach can help users: "I know you probably have a busy schedule. I still think you can manage and hit your goal of daily step count." "Looks like you are having a busy day. I would recommend setting up gentle reminders daily of your goal to have them as part of each day. Hope that can help you be all set for having an exercise routine!" "You know, building a new habit is really really hard. But it doesn't have to be that way :) Starting with a little stroll outside for some fresh air cannot be bad idea as long as the weather is right. So why not head out today for a few minutes, and come in. What do you think? :) " "You must keep that fire burning, your excitement and confidence for maintaining a healthy lifestyle will take you far with healthy habit formation. I believe a daily stroll with be no problem for you at all:)" "So do you reckon you'll manage your walk today?" "It is nice and bright outside today. What is your plan for the day, why not start walking today?" "The question you can ask yourself is that do you feel walking can help you?" "You knew starting a healthy habit can be hard, but it's a life changing experience of rebuilding your identity as someone who exercises :) If you're not feeling up for a long walk today, perhaps we can aim for a shorter one? :)" "You know walking can be especially enjoyable as it allows you to put on your favourite playlist and podcast. So, what do feel like listening to today?" **...** **Coach prompt: Hey John, It's time for your morning walk.**

2.5. Constraining LLM responses using a COM-B classifier

In order to further constrain or guide the LLM to provide nudges based on COM-B principles, we trained two classifiers to assess C/O/M levels:

1. User statement classifier: Given a user statement sentence, the user-query classifier assigns a high vs low value for each of the capability, opportunity and motivation(COM) dimensions (multi-label classification).

2. Coach statement classifier: Given a shortlist of 15 top LLM outputs, the coach response classifier maps each response to either C, O or M (multi-class classification).

The classifiers were designed as follows. The input string (could be multiple sentences) was embedded using a BERT-base model with the final layer finetuned over either a multi-label head (user statement classifier) or 3 separate C/O/M heads (coach statement classifier). Models were optimised with a cross-entropy loss. Separate user and coach classifiers were trained using samples of 432 user statements and 531 coach statements from the PACE study, manually annotated with C/O/M status. These datasets were split 70:10:20 across train, validation and test splits. Weights were not shared between the user and coach models.

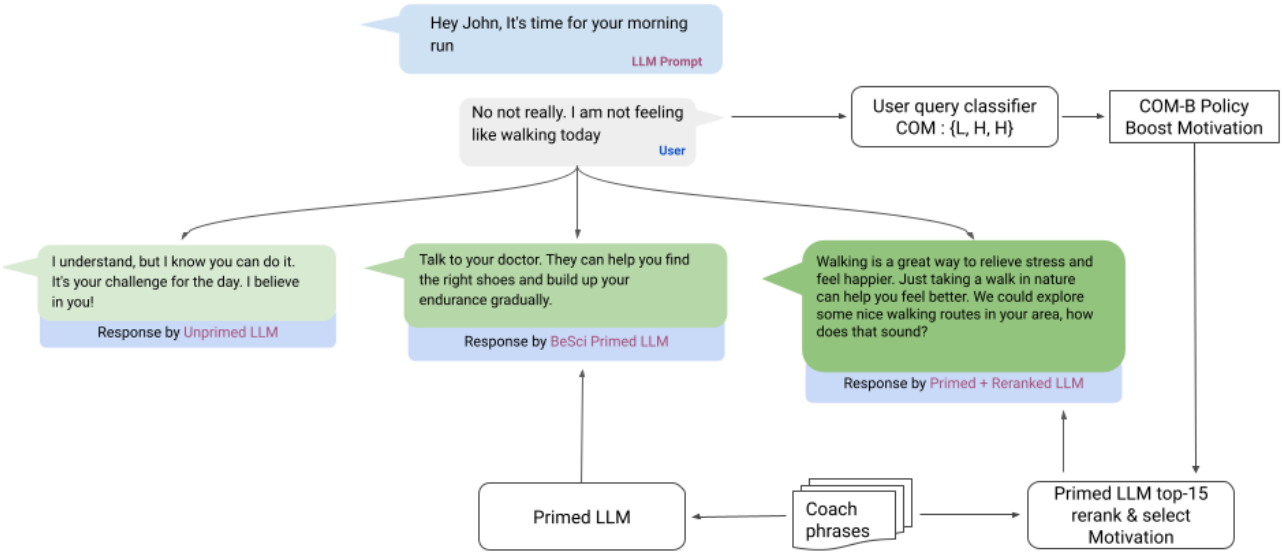


Figure 1: Comparison of example conversations with unprimed, coach-primed and primed+reranked LLMs.

2.6. Simulated dialogue with re-ranking

The simulated conversation experiment was repeated with the primed LaMDA model, using the above classifiers to align the coach response to the inferred user need. For the 9 coach-primed LaMDA transcripts above, a single user statement was manually selected as the most representative of the user's behavioral need.

The selected text was input into the user statement classifier to identify the C/O/M need. The same user text was input into the coach-primed LaMDA model to generate the top 15 candidate responses. These 15 responses were then separately run through the coach statement classifier to generate a likelihood score across each C/O/M category. The coach action was aligned based on the user's inferred C/O/M need based on the rules in Table 2 (i.e. the statement with the highest score in the desired coach action was chosen).

In addition, we conducted an 'oracle' experiment where the user response was manually categorized into C/O/M need and the corresponding coach-primed output was chosen.

Two manual review exercises were then conducted:

1. Comparing the coach-primed output to the classifier re-ranked output; and
2. Comparing coach-primed with the oracle re-ranked output. Note that in both these review exercise, only a single coach response was being adjudicated rather than an entire conversation as previous.

2.7. Evaluation attributes

An evaluation framework was defined based on four key attributes of an LLM-based fitness coach: actionability, realism, motivation and empathy. Coupled with a global assessment of coaching quality, these attributes

Table 2: COM-B policy to select nudge theme based on C/O/M values derived from the user statement classifier.

Capability	Opportunity	Motivation	COM-B Action
Low	High/Low	High/Low	Boost Capability
High	Low	High/Low	Boost Opportunity
High	High	High/Low	Boost Motivation

Table 3: Evaluation attributes cross-referenced with established attributes of coaches and LLMs.

Evaluation attributes	**Coach attributes**	**LLM attributes**
Actionability	Professional competence	Informativeness
Realism	Context sensitivity	Sensibleness & safety
Motivation	BS interventions	Interestingness
Empathy	Social-emotional competences	Groundedness

informed the design of the quantitative and qualitative review detailed below. Table 3) shows how these attributes align with published evaluation frameworks for coaches (40) and for LLMs (21; 41).

2.8. Quantitative review

For each architecture, the unprimed versus coach-primed transcripts generated from the same starting prompt were compared in a pairwise manner. The conversations were evaluated based on the following quantitative attributes: average length of reply, number of conversational turns, user sentiment at conversation end, presence of questions in the coach dialogue and use of coaching-specific words ('goal', 'health', 'routine', 'recover, 'challenge', 'workout', 'training', 'rest'). The results for unprimed versus primed LLMs were compared using a two sided t-test.

2.9. Qualitative review

8 independent reviewers were selected to adjudicate the transcripts. Reviewers were blinded to the manner of LLM priming (naive vs BS) and Re-Ranked LLM variations (naive vs primed vs re-ranked LLM). Raters completed a structured survey with Likert scale responses for the same pairwise comparisons of naive-primed and coach-primed transcripts as above. Questions evaluated the following attributes of the conversation: actionability, realism, empathy, motivation, overall quality. The questions are included in Supplementary Table 2.

Table 4: Quantitative assessment of unprimed versus coach-primed LLM conversations.

Metric	**Unprimed**	**Coach-primed**	**p value**
Average length of LLM reply (# words ± S.D.)	25.7 ± 6.5	23.7 ± 7.1	0.3
Turns of conversation by user/LLM (# turns ± S.D.)	3.1 ± 0.3	3.7 ± 0.7	0.2
Conversations ending with positive user sentiment (%)	30	60	**0.02**
Conversations containing a question asked by LLM (%)	0	30	**0.03**
Conversations containing coaching-specific words used by LLM (%)	40	80	0.08

Table 5: Qualitative assessment of unprimed versus coach-primed LLM conversations based on the reviews of 8 adjudicators.

Survey question (1, strong disagree 5, strong agree)	**Unprimed**	**Coach-primed**	**p value**
Which conversation provides a better overall coaching experience (%, remainder unsure)	21	72	-
The quality of the coaching experience is high	3.42 ± 0.88	3.97 ± 1.0	**<0.001**
The coach provides concrete fitness strategies that are actionable to the user	3.61 ± 1.1	4.25 ± 0.84	**<0.001**
The coach provides motivation or encouragement to the user	3.68 ± 1.1	3.97 ± 0.97	0.095
The coach is empathetic toward the user's needs and challenges	3.51 ± 1.0	3.83 ± 1.1	0.123
The language used by the coach is realistic and appropriate for the setting	3.71 ± 1.1	4.10 ± 1.0	**0.02**

Table 6: Class balance and model performance on C/O/M classification for user statements.

	Class balance (high:low)		Classifier performance (ROC-AUC)
	Train	Test	
Motivation	220:40	68:16	0.86
Capability	158:66	51:40	0.77
Opportunity	112:34	52:13	0.83

3. Results

Quantitative analysis (Table 4) showed that the number of turns of dialogue was higher in coach-primed versus unprimed. Across both architectures, priming was associated with a significant boost in the rate of conversations ending in a positive user sentiment, the rate of question-asking by the coach LLM, and the use of coaching-related vocabulary.

To determine whether ratings for the primed and unprimed models differed from each other, we ran a series of linear mixed model analyses. These included a fixed effect for primed vs unprimed, and random effects for rater and prompt to account for non-independence of the observations. Regarding message content, the ratings of blinded reviewers were overall more favorable for the coach-primed LLMs. The (Table 5). Specifically, the coach-primed model was rated as significantly higher in terms of quality, providing actionable suggestions, and using realistic language. The ratings for the classifier re-ranked versus unprimed were less conclusive, but this may be because those ratings were based on a single statement response from the model rather than a full back-and-forth dialogue.

Tables 6 and 7 show the performance of the user and coach statement classifiers, including the size and label distribution in the train and test sets. The BERT-base model had 81% multi-class accuracy in accurately classifying the coach message as motivation, capability or opportunity.

To quantitatively evaluate the re-ranked response compared to the default response, 8 independent reviewers rated both the responses across several dimensions of activity coaching (Table 8). Based on Likert scale responses, the re-ranked answers were rated as more actionable [3.66±0.89 vs 2.88±0.85]; however the other attributes did not reach statistical significance.

Table 7: Model performance on C/O/M classification for coach statements.

Category	Train	Test	Precision	Recall	F1 Score	Multi-class accuracy
Motivation	256	121	0.87	0.86	0.86	0.81
Capability	139	66	0.88	0.72	0.79	
Opportunity	212	74	0.83	0.71	0.77	

Table 8: Qualitative review of coach-primed versus classifier re-ranked and oracle re-ranked dialogues.

Survey question (1, strong disagree → 5, strong agree)	Unprimed	Coach-primed	Classifier Re-ranked	Oracle Re-ranked	p value (Classifier re-ranked vs Unprimed
The coach response provided concrete fitness strategies that are actionable	2.88 ± 0.85	3.22 ± 0.95	3.66 ± 0.89	4.02 ± 0.68	**0.02**
The coach response to user questions was in a realistic manner	3.02 ± 0.98	3.23 ± 0.97	3.59 ± 0.99	3.75 ± 0.80	0.18
The coach response provided motivation or encouragement to the user	3.05 ± 1.0	3.19 ± 0.85	3.75 ± 0.92	3.45 ± 1.05	0.36
The coach is empathetic toward the user's needs and challenges	2.94 ± 0.99	3.05 ± 0.94	3.56 ± 0.87	3.77 ± 0.83	0.47
The language used by the coach is realistic and appropriate for the setting	3.33 ± 0.87	3.48 ± 0.84	3.69 ± 0.87	3.78 ± 0.79	0.29
Average total score	3.04 ± 0.95	3.24 ± 1.36	3.65 ± 1.32	3.75 ± 0.84	

4. Discussion

This proof-of-concept study introduces two methods to infuse behavior science into LLM dialogue. We demonstrate that behavior science-based priming is a simple but effective strategy to tailor LLMs for activity coaching, with specific benefits in terms of actionability and the provision of concrete coaching advice. Additionally, post-hoc re-ranking of LLM responses based on behavior science principles can further enhance attributes such as perceived empathy.

BS priming yielded some significant boosts in various proxies for coaching quality. This trend was evident across both quantitative and qualitative metrics. Notably, coach phrase priming was associated with a higher number of conversational turns, a greater rate of question-asking, and more frequent use of coaching vocabulary. Manual review also judged coach phrase priming as providing significantly greater motivation and concrete coaching strategies versus the unprimed LLM. This suggests that BS priming may be an effective and accessible strategy for customising LLMs for various coaching scenarios.

A unique aspect of this work is the combination of priming with post-hoc re-ranking to enable knowledge infusion at multiple touchpoints. Interestingly, re-ranking resulted in significant incremental improvements in actionability, with upward trends in empathy, motivation and realism that did not meet statistical significance. We demonstrate this uplift both for a classifier-based re-ranking, which introduces error from mis-classification; and for oracle-based re-ranking, which showed a further marginal advantage over the former. Together, these results demonstrate the ability to stitch together multiple simple constraints as part of a hybrid knowledge infusion strategy. As LLMs become more pervasive in the coaching domain, this will be increasingly important.

Since Capability has marginally lower user statement classifier accuracy, it was wrongly identified as motivation in few cases of classifier based BeSci dialogue alignment LLM. This resulted in higher motivational character to classifier based LLM over Oracle LLM at the expense of lower empathy and actionability scores.

This study has a number of limitations. First, the evaluation was predominantly based on simulated conversations with a single human interacting with the LLMs, which invariably introduces bias even in the presence of blinding. Future work could trial a similar evaluation with larger groups of users engaging in dialogue, as per [ref]. The rudimentary priming method used here could be extended, e.g. by more explicit instruction prompting or chain of thought prompting. The re-ranking method was limited in only focusing on a single user query and coach response. In reality, it is important to consistently align the coach responses to user need throughout a conversation and adapt as the dialogue unfolds. Methods such as reinforcement learning with human feedback can help to offer this adaptability (29). Finally, the behaviour model used was a simplistic one that conceptualizes user behaviour only along three axes - future studies could consider using more sophisticated behavior science frameworks, which may help to better target coach actions.

5. Conclusion

Knowledge infusion methods based on behavior science principles can be used to improve the quality of LLM-generated physical activity related conversations. The combination of coach phrase priming with re-ranking of LLM outputs offers optimal results in terms of manually-adjudicated actionability, empathy and overall coaching experience. These methods can help to constrain and guide LLMs in various coaching scenarios.

References

[1] Regina Guthold, Gretchen A Stevens, Leanne M Riley, and Fiona C Bull. Worldwide trends in insufficient physical activity from 2001 to 2016: a pooled analysis of 358 population-based surveys with 1·9 million participants. *Lancet Glob Health*, 6(10):e1077–e1086, October 2018.

[2] I-Min Lee, Eric J Shiroma, Felipe Lobelo, Pekka Puska, Steven N Blair, Peter T Katzmarzyk, and Lancet Physical Activity Series Working Group. Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. *Lancet*, 380(9838):219–229, July 2012.

[3] WHO. GLOBAL ACTION PLAN ON PHYSICAL ACTIVITY 2018-2030: More active people for a healthier world. Technical report, World Health Organization, 2018.

[4] Judit Bort-Roig, Nicholas D Gilson, Anna Puig-Ribera, Ruth S Contreras, and Stewart G Trost. Measuring and influencing physical activity with smartphone technology: a systematic review. *Sports Med.*, 44(5):671–686, May 2014.

[5] Alexandher Negreiros, Roberto B T Maciel, Bianca Carvalho de Barros, and Rosimeire Simprini Padula. Quality assessment of smartphone fitness apps used to increase physical activity level and improve general health in adults: A systematic review. *Digit Health*, 8:20552076221138305, November 2022.

[6] Art of the nudge. https://www.omadahealth.com/art-of-the-nudge. Accessed: 2023-7-3.

[7] Anouk Middelweerd, Julia S Mollee, C Natalie van der Wal, Johannes Brug, and Saskia J Te Velde. Apps to promote physical activity among adults: a review and content analysis. *Int. J. Behav. Nutr. Phys. Act.*, 11:97, July 2014.

[8] Ashish Chaddha, Elizabeth A Jackson, Caroline R Richardson, and Barry A Franklin. Technology to help promote physical activity. *Am. J. Cardiol.*, 119(1):149–152, January 2017.

[9] Gemma Flores Mateo, Esther Granado-Font, Carme Ferré-Grau, and Xavier Montaña-Carreras. Mobile phone apps to promote weight loss and increase physical activity: A systematic review and Meta-Analysis. *J. Med. Internet Res.*, 17(11):e253, November 2015.

[10] Timothy W Bickmore, Rebecca A Silliman, Kerrie Nelson, Debbie M Cheng, Michael Winter, Lori Henault, and Michael K Paasche-Orlow. A randomized controlled trial of an automated exercise coach for older adults. *J. Am. Geriatr. Soc.*, 61(10):1676–1683, October 2013.

[11] Merav Allouch, Amos Azaria, and Rina Azoulay. Conversational agents: Goals, technologies, vision and challenges. *Sensors*, 21(24), December 2021.

[12] Ahmet Baki Kocaballi, Shlomo Berkovsky, Juan C Quiroz, Liliana Laranjo, Huong Ly Tong, Dana Rezazadegan, Agustina Briatore, and Enrico Coiera. The personalization of conversational agents in health care: Systematic review. *J. Med. Internet Res.*, 21(11):e15360, November 2019.

[13] Mira El Kamali, Leonardo Angelini, Maurizio Caon, Giuseppe Andreoni, Omar Abou Khaled, and Elena Mugellini. Towards the NESTORE e-coach: a tangible and embodied conversational agent for older adults. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, UbiComp '18, pages 1656–1663, New York, NY, USA, October 2018. Association for Computing Machinery.

[14] Genta Indra Winata, Holy Lovenia, Etsuko Ishii, Farhad Bin Siddique, Yongsheng Yang, and Pascale Fung. Nora: The Well-Being coach. June 2021.

[15] Rafal Kocielnik, Lillian Xiao, Daniel Avrahami, and Gary Hsieh. Reflection companion: A conversational system for engaging users in reflection on physical activity. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(2):1–26, July 2018.

[16] Madhurima Vardhan, Narayan Hegde, Srujana Merugu, Shantanu Prabhat, Deepak Nathani, Martin Seneviratne, Nur Muhammad, Pranay Reddy, Sriram Lakshminarasimhan, Rahul Singh, Karina Lorenzana, Eshan Motwani, Partha Talukdar, and Aravindan Raghuveer. Walking with PACE - personalized and automated coaching engine. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '22, pages 57–68, New York, NY, USA, July 2022. Association for Computing Machinery.

[17] Jingwen Zhang, Yoo Jung Oh, Patrick Lange, Zhou Yu, and Yoshimi Fukuoka. Artificial intelligence chatbot behavior change model for designing artificial intelligence chatbots to promote physical activity and a healthy diet: Viewpoint. *J. Med. Internet Res.*, 22(9):e22845, September 2020.

[18] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are Few-Shot learners. May 2020.

[19] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. PaLM: Scaling language modeling with pathways. April 2022.

[20] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling language models: Methods, analysis & insights from training gopher. December 2021.

[21] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, Yaguang Li, Hongrae Lee, Huaixiu Steven Zheng,

Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. LaMDA: Language models for dialog applications. January 2022.

[22] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief. Bioinform.*, 23(6), November 2022.

[23] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. 2022.

[24] Adam Sobieszek and Tadeusz Price. Playing games with ais: The limits of GPT-3 and similar large language models. *Minds Mach.*, 32(2):341–364, June 2022.

[25] Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[26] Yihan Wang, Si Si, Daliang Li, Michal Lukasik, Felix Yu, Cho-Jui Hsieh, Inderjit S Dhillon, and Sanjiv Kumar. Preserving In-Context learning ability in large language model fine-tuning. November 2022.

[27] Fedor Moiseev, Zhe Dong, Enrique Alfonseca, and Martin Jaggi. SKILL: Structured knowledge infusion for large language models. May 2022.

[28] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. April 2022.

[29] Banghua Zhu, Jiantao Jiao, and Michael I. Jordan. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons, 2023.

[30] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for Parameter-Efficient prompt tuning. April 2021.

[31] Yanchen Liu, Timo Schick, and Hinrich Schütze. Semantic-Oriented unlabeled priming for Large-Scale language models. February 2022.

[32] Simran Arora, Avanika Narayan, Mayee F Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami, Frederic Sala, and Christopher Ré. Ask me anything: A simple strategy for prompting language models. October 2022.

[33] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are Human-Level prompt engineers. November 2022.

[34] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. AI chains: Transparent and controllable Human-AI interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, number Article 385 in CHI '22, pages 1–22, New York, NY, USA, April 2022. Association for Computing Machinery.

[35] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-Most prompting enables complex reasoning in large language models. May 2022.

[36] Jayr Pereira, Robson Fidalgo, Roberto Lotufo, and Rodrigo Nogueira. Visconde: Multi-document QA with GPT-3 and neural reranking. December 2022.

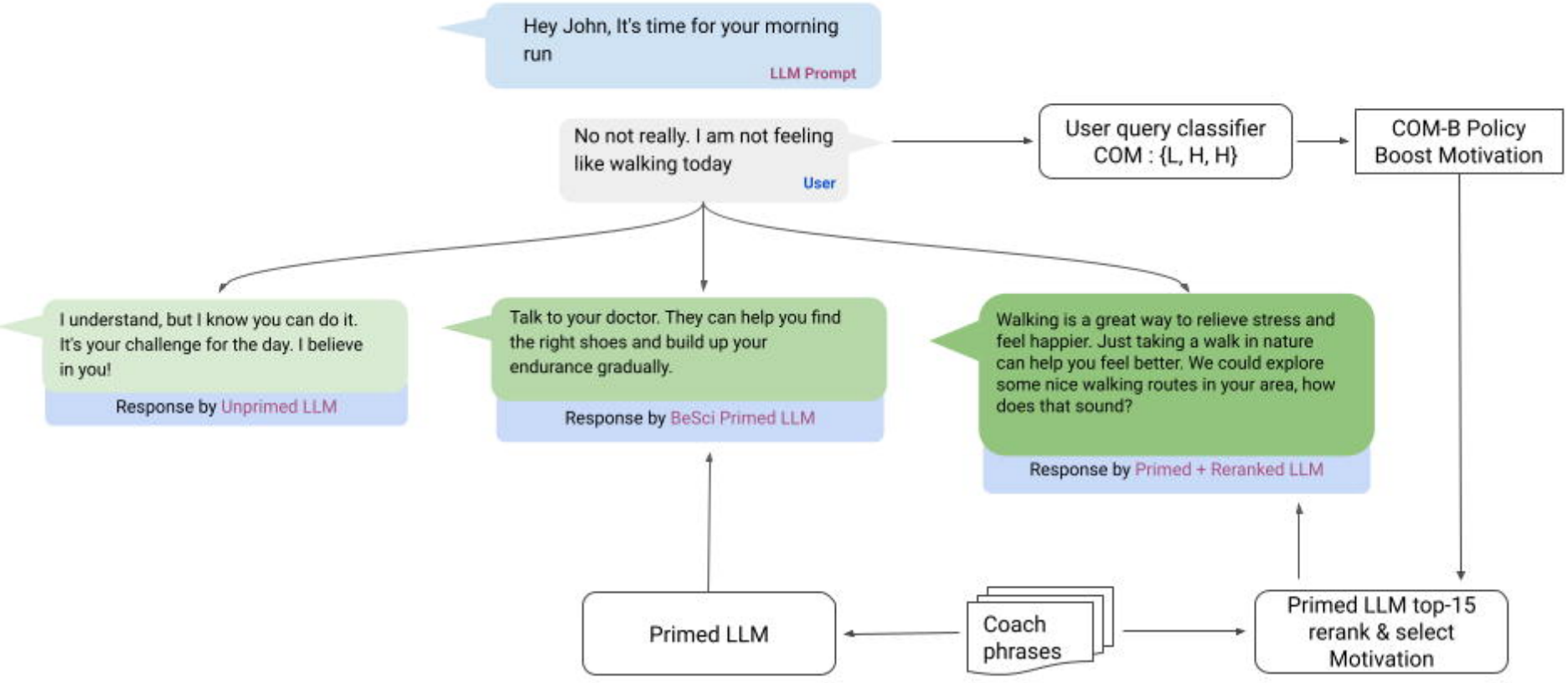
[37] Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. Prompt-and-Rerank: A method for Zero-Shot and Few-Shot arbitrary textual style transfer with small language models. May 2022.

[38] Susan Michie, Maartje M van Stralen, and Robert West. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. *Implement. Sci.*, 6:42, April 2011.

[39] Aditya Kumar Purohit, Louis Barclay, and Adrian Holzer. Designing for digital detox: Making social media less addictive with digital nudges. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, pages 1–9, New York, NY, USA, April 2020. Association for Computing Machinery.

[40] Ulrich Georg Strauch, Hagen Wäsche, and Darko Jekauc. Coach competences to induce positive affective reactions in sport and Exercise-A qualitative study. *Sports (Basel)*, 7(1), January 2019.

[41] Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V Le. Towards a human-like Open-Domain chatbot. January 2020.


Hey John, It's time for your morning run
LLM Prompt
No not really. I am not feeling like walking today
User
User query classifier COM : {L, H, H}
COM-B Policy Boost Motivation
I understand, but I know you can do it. It's your challenge for the day. I believe in you!
Response by Unprimed LLM
Talk to your doctor. They can help you find the right shoes and build up your endurance gradually.
Response by BeSci Primed LLM
Walking is a great way to relieve stress and feel happier. Just taking a walk in nature can help you feel better. We could explore some nice walking routes in your area, how does that sound?
Response by Primed + Reranked LLM
Primed LLM
Coach phrases
Primed LLM top-15 rerank & select Motivation