Priming Large Language Models for Personalized Healthcare

Madhurima Vardhan
Argonne Leadership Computing Facility
Argonne National Laboratory
Lemont, USA
0000-0003-4019-7832

Deepak Nathani
Department of Computer Science
University of California, Santa Barbara
Santa Barbara, USA

Swarnima Vardhan
Department of Internal Medicine
Yale New Haven Health, Bridgeport Hospital
Bridgeport, USA

Abhinav Aggarwal
Department of Internal Medicine
Yale New Haven Health, Bridgeport Hospital
Bridgeport, USA

***Abstract*—Large Language Models (LLMs) have captured attention of researchers across different scientific fields. However, sensitive data access issues, model retraining, long compute time and lack of real-time results have limited the direct application of LLMs in fields such as healthcare fitness. Healthcare fitness is ripe to take advantage of the near-human efficiency and accuracy of LLMs due to ever increasing gap between human coaches and population that requires fitness coaching. In this work, we introduce a lightweight approach, priming LLMs, to develop an automated health coach that relies upon fundamental theories of behavior science and taps into the enormous potential of LLMs. We found that sentence length and conversation length were higher in primed LLMs compared to naïve context aware LLMs. Subsequently, we conducted a qualitative reviewer evaluation and report that the primed architectures were overall more appropriate and demonstrated higher empathy.**

***Index Terms*—Large Language Models, Personalized healthcare, fitness coaching, prompt engineering**

I. Introduction

Automated and personalized health coach assistants have the potential to reduce the cost of fitness and need for trained coaches who are required to cater the ever-increasing population suffering from non-communicable diseases and the rampant sedentary lifestyles [1] [2]. Given the rise of interest in personal health monitoring systems and increasing disparity with respect to the number of trained coaches versus the number of people who require fitness coaching, Large Language Models (LLMs) can be offered as an attractive solution to function as an automated health assistant. LLMs offer flexibility in performing a series of generalized tasks with near-human efficiency and accuracy [3]. LLMs such as GPT-3, have shown promise for task-oriented dialogue across a range of domains [3]. Both LaMDA and GPT-3 use the Transformer-based neural language models specialized for dialog applications [3]. In this work, we introduce a lightweight approach that constrains generalized LLMs to the specific task of functioning as a fitness coach and relies on established behavioral science models to enable empathetic and personalized conversations under different coaching scenarios.

While adapting or post-training LLMs using an unlabeled domain corpus has the potential to improve performance for end-tasks in a particular domain, the limitations around access to healthcare and personal data impede the application of LLMs for developing a personalized automated conversational assistant for fitness coaching [4] [5]. Thus, the use of LLMs in exercise coaching conversations remains relatively under-explored. Yet another reason for the lack of real-world automated fitness assistants using LLMs is also in part due to the complexity associated with health behavior change [6]. The field of behavior science has developed numerous frameworks for analyzing and influencing user behaviors, which has been critical in the design of personalized nudging programs in healthcare and fitness [7]. One such model is the Fogg's Behavior Model (FBM) that asserts the target behavior change of user can be explained across three axes by assessing: (1) motivation – is the user sufficiently motivated (2) ability – does the user have the ability to perform the given task, and (3) propensity – can the user plan or be triggered to perform the target activity. Several automated health assistant frameworks using different machine learning models have relied on the application of FBM to target behavior change, specifically for fitness coaching [8].

In this work, we explore how behavior science models, such as the FBM, could be infused into an LLM, and be used to constrain and/or guide the coaching conversations in a way that is consistent with established practice of human coaches. Towards this end, we propose priming LLMs as a lightweight approach that does not require additional model retraining and therefore precludes the need for any prior coach-user conversations. Priming essentially comprises of prompt engineering and design that can allow the model to be constrained for a specific task and in return has a higher probability to generate a more appropriate, favorable, and contextual conversation [9]. We encapsulate the FBM by priming the LLMs with example coach responses, mapped to the motivation, ability, and propensity of a user. Subsequently,

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

we qualitatively assess the conversations by independent human reviewers (n=5) generated by primed and unprimed LLMs across a commonly used LLM architectures, GPT-3. We found that sentence length, conversation length was higher in primed LLMs. Reviewers found that the primed architectures were overall more appropriate, showed higher motivation and greater empathy. Together, these experiments serve as a proof of concept of how LLMs and behavior science might be integrated, laying the foundation for future work around knowledge infusion in these conversational agents.

In the remainder sections of this work, Section II describes the overall approach and in Section III we discuss preliminary results, conclusion and future work.

II. Methods

In this section we describe our approach of priming LLMs with the FBM. ***Incorporating behavior science in LLMs by priming.*** We created a repository consisting of coach responses to different user scenarios by consulting expert behavior scientists and trained fitness coaches. The curated coach responses were tailored to a specific coach action across the three axes of FBM. For example, in the context of motivation, appropriate coach responses were created for encouragement, fun/temptation bundling, congratulating, and exemplifying core values/perceived benefit. Consequently, for ascertaining user ability, coach responses were constructed around providing educational information, barrier conversations, and recovery. Also, for propensity, appropriate examples were created for having goal conversations and activity planning. Using these examples, we primed GPT-3, and we refer to this approach as Behavior Science-based priming. We set the following model parameters for the open-source GPT-3 model, temperature (controls model randomness) to 0.9, maximum token length of 1024, top P (controls response diversity for likelihood responses) to 1 and frequency penalty (probability of verbatim model responses) to 0.9 and presence penalty (controls likelihood for new topics) to 0.6. We measure the quality of coach responses of the behavior science (BS) primed model, by comparing to naïve context-aware LLM model. For both the models (naïve and BS-primed), we emulated 10 different user scenarios exemplifying the need to elicit coach actions for sensing and boosting motivation, ability, and propensity of a user in a real-world scenario. We qualitatively evaluated the coach responses from both BS-primed and naïve LLMs by asking independent reviewers (n=5) to rate the conversations along different conversation dimensions, for example, coaching experience, empathy and appropriateness.

III. Conclusion and Future Work

Preliminary Results. Evaluating Behavior science (BS) primed LLM and naïve context aware LLM to function as an automated fitness coach. Qualitative analysis of coach-user conversations for both the BS-primed and naïve context aware, revealed that coach actions along the three FBM axes of motivation, ability and propensity were well-represented. However, we found that sentence length and conversation length were higher in BS-primed LLMs compared to naïve context aware LLMs. Furthermore, we qualitatively evaluated the quality of conversations of BS-primed and naïve context aware LLMs and found that all reviewers (n=5) preferred the BS-primed LLM responses with respect to coaching experience, empathy and appropriateness.

Conclusion and Future Work This is a proof of concept study of how fundamental models and medical knowledge can be used to encode healthcare information in LLM and enable them to function as an automated medical assistants for a more personalized experience without the need for any additional model retraining. User data such as obtained from wearable devices and smartphones can be used for automating the prompts via priming. Based on this framework, we will develop a zero shot learning approach for priming a LLMs that can function as an automated medical assistant for different clinical tasks, such that users will be able to directly chat with the LLM. Furthermore, we will quantitatively evaluate these tasks by human raters having domain expertise. We expect the ratings for the primed LLM to be significantly higher in terms of domain knowledge and empathy. As part of our future efforts, we will compare experimental results to real coaching assistant and automated virtual assistants.

IV. Acknowledgment

This research used resources of the Argonne Leadership Computing Facility, which is a U.S. Department of Energy Office of Science User Facility operated under contract DE-AC02-06CH11357.

References

[1] K. Matthias et al., "The mobile fitness coach: Towards individualized skill assessment using personalized mobile devices," Pervasive and Mobile Computing, vol. 9, no. 2, pp. 203-215, 2013.

[2] M. Vardhan et al.. "Walking with PACE-Personalized and Automated Coaching Engine." In Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization, pp. 57-68. 2022.

[3] T. Brown et al., "Language models are few-shot learners." Advances in neural information processing systems 33: 1877-1901., 2020,

[4] Z. Ke et al., "Continual training of language models for few-shot learning." arXiv preprint arXiv:2210.05549, 2022.

[5] S. Yeom et al., "Privacy risk in machine learning: Analyzing the connection to overfitting." In 2018 IEEE 31st computer security foundations symposium (CSF), pp. 268-282. IEEE, 2018.

[6] S. Michie et al., "The Human Behaviour-Change Project: harnessing the power of artificial intelligence and machine learning for evidence synthesis and interpretation." Implementation Science 12, no. 1: 1-12, 2017.

[7] B. J. Fogg, "A behavior model for persuasive design." In Proceedings of the 4th international Conference on Persuasive Technology, pp. 1-7. 2009.

[8] A. Lisowska et al., "From Personalized Timely Notification to Healthy Habit Formation: a Feasibility Study of Reinforcement Learning Approaches on Synthetic Data." In SMARTERCARE@ AI* IA, pp. 7-18. 2021.

[9] L. Reynolds et al., "Prompt programming for Large Language Models: Beyond the few-shot paradigm." In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1-7. 2021.

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

Copyright (c) IARIA, 2024. ISBN: 978-1-68558-136-7