Electronic Cover Sheet

PI: **Rogers, Ethan Barrett**	Title: Building and evaluating a scalable health coaching system with large language models to promote physical activity in older adults	
Received: 08/08/2024	Opportunity: PA-23-272	Council: 01/2025
Competition ID: FORMS-H	FOA Title: Ruth L. Kirschstein National Research Service Award (NRSA) Individual Predoctoral Fellowship (Parent F31)	
1F31HL176319-01A1	Dual: AG	Accession Number: 5035652
IPF: 6116101	Organization: NORTHEASTERN UNIVERSITY	
Former Number: 1F31HL176319-01	Department:	
IRG/SRG: ZRG1 F16-K (20)L	AIDS: N	Expedited: N
Subtotal Direct Costs (excludes consortium F&A)	Animals: N Humans: Y Clinical Trial: N Current HS Code: XM HESC: N HFT: N	New Investigator: Early Stage Investigator:

Senior/Key Personnel:	*Organization:*	*Role Category:*
Ethan Rogers	NORTHEASTERN UNIVERSITY	PD/PI
Dakuo Wang	Northeastern University	Other (Specify)-co-sponsor
STEPHEN INTILLE	Northeastern University	Other (Specify)-co-sponsor
Dinesh John	Northeastern University	Other (Specify)-sponsor

Reference Letters

[redacted]

424 R&R and PHS-398 Specific
Table Of Contents

OMB Number: 4040-0010
Expiration Date: 11/30/2025

Project/Performance Site Location(s)

Project/Performance Site Primary Location

❍ I am submitting an application as an individual, and not on behalf of a company, state, local or tribal government, academia, or other type of organization.

Organization Name:	NORTHEASTERN UNIVERSITY
UEI:	HLTMVS2JZBS6
Street1*:	NORTHEASTERN UNIVERSITY
Street2:	360 HUNTINGTON AVE, 177-500
City*:	BOSTON
County:	
State*:	MA: Massachusetts
Province:	
Country*:	USA: UNITED STATES
Zip / Postal Code*:	021155005
Project/Performance Site Congressional District*:	ma-007

Additional Location(s) File Name:

RESEARCH & RELATED Other Project Information

1. Are Human Subjects Involved?* ● **Yes** ❍ **No**

1.a. If YES to Human Subjects

Is the Project Exempt from Federal regulations? ● Yes ❍ No

If YES, check appropriate exemption number: 1 ✔ 2 3 ✔ 4 5 6 7 8

If NO, is the IRB review Pending? ❍ **Yes** ❍ **No**

IRB Approval Date:

Human Subject Assurance Number [redacted]

2. Are Vertebrate Animals Used?* ❍ **Yes** ● **No**

2.a. If YES to Vertebrate Animals

Is the IACUC review Pending? ❍ Yes ❍ No

IACUC Approval Date:

Animal Welfare Assurance Number

3. Is proprietary/privileged information included in the application?* ❍ **Yes** ● **No**

4.a. Does this project have an actual or potential impact - positive or negative - on the environment?* ❍ **Yes** ● **No**

4.b. If yes, please explain:

4.c. If this project has an actual or potential impact on the environment, has an exemption been authorized or an environmental assessment (EA) or environmental impact statement (EIS) been performed? ❍ Yes ❍ No

4.d. If yes, please explain:

5. Is the research performance site designated, or eligible to be designated, as a historic place?* ❍ **Yes** ● **No**

5.a. If yes, please explain:

6. Does this project involve activities outside the United States or partnership with international collaborators?* ❍ **Yes** ● **No**

6.a. If yes, identify countries:

6.b. Optional Explanation:

	Filename
7. Project Summary/Abstract*	Abstract_RESUB_FINAL.pdf
8. Project Narrative*	Narritive_RESUB_FINAL.pdf
9. Bibliography & References Cited	Bibliography_RESUB_FINAL.pdf
10.Facilities & Other Resources	Facilities_and_Rescources_RESUB_FINAL.pdf
11.Equipment	Equipment_RESUB_FINAL.pdf

Abstract

Physical activity prevents and treates many types of chronic illness, such as Altzheimer's Disease and orther neurodegenerative illness. Older adults (>60) face the greatest chronic disease burden, thus requiring interventiosn tailored to their needs. Traditional physical activity health coaching (PAHC) interventions, while an effective to increase PA in older adults, face scalability challenges due to the time and effort required by health coaches to establish and maintain meaningful relationships with clients. To expand the reach and productivity of health coaches, I propose a novel Human-in-the-loop Artificial Intelligence (AI) PAHC system, wherein human health coaches are augmented with responses from AI. Previous approaches to AI PAHC in older adults have had limited success, with lack of access to large PAHC-focused training datasets and lack of personalization and adaptability cited as key limitations. I will address these barriers to success by using a Large Language Model (LLM) fine-tuned with 20k PAHC messages between health coaches and older adults, embedded in a LLM-Augmenter framework to integreate external knowledge and reduce factual inaccuracy. LLMs exhibit advanced capabilities in natural language understanding and the ability to meaningfully integrate personal details into responses but are also liable to produce factual inaccuracies in their responses, deemed 'hallucinations.' The LLM-Augmenter framework has been shown to eliminate these hallucinations by integreating external knowledge into the model's responses. We hypothesize that fine-tuning an LLM with the PAHC messages and incorporating an LLM-Augmenter to query external knowledge sets and mitigate hallucination will yield an AI model capable of replicating the response quality of human health coaches given the same input. In **Aim 1**, I outline the construction of this model and the development of a user interface for health coaches to message participants with AI-generated responses. **Aim 2** details the evaluation of AI PAHC responses using metric-based testing and human evaluation of the systems responses and user interface. BERTscore, a metric capturing intent and meaning similarity, will assess how closely our AI replicates human health coach responses. Additionally, health coachs (N=5) will rate the humanness, adherence to theoretical frameworks, and degree of hallucination present in the system response. Health coaches will use the interface and provide qualitative insights, ensuring a comprehensive evaluation. Future research will focus on evaluation of the system's ability to increase health coach productivity, as well as futher automation of the PAHC process.

Narritive

Physical Activity health coaching (PAHC) effectively prevents and treats chronic illnesses in older adults but is difficult to scale due to its reliance on human health coaches. By leveraging new AI technologies such as large language models (LLMs) and a PAHC message set, an AI PAHC system will be developed and evaluated to work alongside human coaches, enhancing their reach and productivity. This innovative approach has the potential to expand access to health coaching for more older adults, addressing the pressing issue of chronic illness while maintaining quality.

BIBLIOGRAPHY

1. U.S. Department of Health and Human Services. *Physical Activity Guidelines for Americans Midcourse Report: Implementation Strategies for Older Adults*. U.S. Department of Health and Human Services; 2023. https://health.gov/sites/default/files/2023-06/PAG_MidcourseReport_508c_final.pdf
2. Boersma P, Black LI, Ward BW. Prevalence of Multiple Chronic Conditions Among US Adults, 2018. *Prev Chronic Dis*. 2020;17:E106. doi:10.5888/pcd17.200130
3. Oliveira JS, Sherrington C, Amorim AB, Dario AB, Tiedemann A. What is the effect of health coaching on physical activity participation in people aged 60 years and over? A systematic review of randomised controlled trials. *Br J Sports Med*. 2017;51(19):1425-1432. doi:10.1136/bjsports-2016-096943
4. An S, Song R. Effects of health coaching on behavioral modification among adults with cardiovascular risk factors: Systematic review and meta-analysis. *Patient Educ Couns*. 2020;103(10):2029-2038. doi:10.1016/j.pec.2020.04.029
5. Bevilacqua R, Casaccia S, Cortellessa G, et al. Coaching Through Technology: A Systematic Review into Efficacy and Effectiveness for the Ageing Population. *Int J Environ Res Public Health*. 2020;17(16):5930. doi:10.3390/ijerph17165930
6. Domingos C, Pêgo JM, Santos NC. Effects of physical activity on brain function and structure in older adults: A systematic review. *Behav Brain Res*. 2021;402:113061. doi:10.1016/j.bbr.2020.113061
7. An HY, Chen W, Wang CW, Yang HF, Huang WT, Fan SY. The Relationships between Physical Activity and Life Satisfaction and Happiness among Young, Middle-Aged, and Older Adults. *Int J Environ Res Public Health*. 2020;17(13):4817. doi:10.3390/ijerph17134817
8. Zhang J, Oh YJ, Lange P, Yu Z, Fukuoka Y. Artificial Intelligence Chatbot Behavior Change Model for Designing Artificial Intelligence Chatbots to Promote Physical Activity and a Healthy Diet: Viewpoint. *J Med Internet Res*. 2020;22(9):e22845. doi:10.2196/22845
9. Mitchell E, Elhadad N, Mamykina L. Examining AI Methods for Micro-Coaching Dialogs. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI ’22. Association for Computing Machinery; 2022:1-24. doi:10.1145/3491102.3501886
10. Arguello D, Rogers E, Denmark GH, et al. Companion: A Pilot Randomized Clinical Trial to Test an Integrated Two-Way Communication and Near-Real-Time Sensing System for Detecting and Modifying Daily Inactivity among Adults >60 Years—Design and Protocol. *Sensors*. 2023;23(4). doi:10.3390/s23042221
11. Peng B, Galley M, He P, et al. Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback. Published online March 8, 2023. doi:10.48550/arXiv.2302.12813
12. Touvron H, Martin L, Stone K, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. Published online July 19, 2023. Accessed July 20, 2023. http://arxiv.org/abs/2307.09288
13. Li Y, Li Z, Zhang K, Dan R, Zhang Y. ChatDoctor: A Medical Chat Model Fine-tuned on LLaMA Model using Medical Domain Knowledge. Published online April 18, 2023. Accessed May 8, 2023. http://arxiv.org/abs/2303.14070
14. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. Bertscore: Evaluating text generation with bert. *ArXiv Prepr ArXiv190409675*. Published online 2019.
15. Durstine JL, Gordon B, Wang Z, Luo X. Chronic disease and the link to physical activity. *J Sport Health Sci*. 2013;2(1):3-11. doi:10.1016/j.jshs.2012.07.009
16. Brown JC, Winters-Stone K, Lee A, Schmitz KH. Cancer, Physical Activity, and Exercise. *Compr Physiol*. 2012;2(4):2775-2809. doi:10.1002/cphy.c120005
17. Yoo TK, Oh BK, Lee MY, Sung KC. Association between physical activity and insulin resistance using the homeostatic model assessment for insulin resistance independent of waist circumference. *Sci Rep*. 2022;12(1):6002. doi:10.1038/s41598-022-10032-2
18. Paffenbarger RS, Hyde RT, Wing AL, Lee IM, Jung DL, Kampert JB. The Association of Changes in Physical-Activity Level and Other Lifestyle Characteristics with Mortality among Men. *N Engl J Med*. 1993;328(8):538-545. doi:10.1056/NEJM199302253280804
19. Shi Y, Li T, Wang Y, et al. Household physical activity and cancer risk: a systematic review and dose-response meta-analysis of epidemiological studies. *Sci Rep*. 2015;5(1):14901. doi:10.1038/srep14901
20. McTiernan A, Friedenreich CM, Katzmarzyk PT, et al. Physical Activity in Cancer Prevention and Survival: A Systematic Review. *Med Sci Sports Exerc*. 2019;51(6):1252-1261. doi:10.1249/MSS.0000000000001937
21. BERLIN JA, COLDITZ GA. A META-ANALYSIS OF PHYSICAL ACTIVITY IN THE PREVENTION OF CORONARY HEART DISEASE. *Am J Epidemiol*. 1990;132(4):612-628. doi:10.1093/oxfordjournals.aje.a115704

22. Lee CD, Folsom AR, Blair SN. Physical Activity and Stroke Risk. *Stroke*. 2003;34(10):2475-2481. doi:10.1161/01.STR.0000091843.02517.9D
23. Aune D, Norat T, Leitzmann M, Tonstad S, Vatten LJ. Physical activity and the risk of type 2 diabetes: a systematic review and dose–response meta-analysis. *Eur J Epidemiol*. 2015;30(7):529-542. doi:10.1007/s10654-015-0056-z
24. Füzéki E, Engeroff T, Banzer W. Health Benefits of Light-Intensity Physical Activity: A Systematic Review of Accelerometer Data of the National Health and Nutrition Examination Survey (NHANES). *Sports Med*. 2017;47(9):1769-1793. doi:10.1007/s40279-017-0724-0
25. Cunningham C, O' Sullivan R, Caserotti P, Tully MA. Consequences of physical inactivity in older adults: A systematic review of reviews and meta-analyses. *Scand J Med Sci Sports*. 2020;30(5):816-827. doi:10.1111/sms.13616
26. Zhu J, Ge F, Zeng Y, et al. Physical and Mental Activity, Disease Susceptibility, and Risk of Dementia. *Neurology*. 2022;99(8):e799-e813. doi:10.1212/WNL.0000000000200701
27. Groot C, Hooghiemstra AM, Raijmakers PGHM, et al. The effect of physical activity on cognitive function in patients with dementia: A meta-analysis of randomized control trials. *Ageing Res Rev*. 2016;25:13-23. doi:10.1016/j.arr.2015.11.005
28. Kivelä K, Elo S, Kyngäs H, Kääriäinen M. The effects of health coaching on adult patients with chronic diseases: A systematic review. *Patient Educ Couns*. 2014;97(2):147-157. doi:10.1016/j.pec.2014.07.026
29. Kreuter MW, Bull FC, Clark EM, Oswald DL. Understanding how people process health information: a comparison of tailored and nontailored weight-loss materials. *Health Psychol Off J Div Health Psychol Am Psychol Assoc*. 1999;18(5):487-494. doi:10.1037//0278-6133.18.5.487
30. Oh YJ, Zhang J, Fang ML, Fukuoka Y. A systematic review of artificial intelligence chatbots for promoting physical activity, healthy diet, and weight loss. *Int J Behav Nutr Phys Act*. 2021;18(1):160. doi:10.1186/s12966-021-01224-6
31. Luo TC, Aguilera A, Lyles CR, Figueroa CA. Promoting Physical Activity Through Conversational Agents: Mixed Methods Systematic Review. *J Med Internet Res*. 2021;23(9):e25486. doi:10.2196/25486
32. Milne-Ives M, Cock C de, Lim E, et al. The Effectiveness of Artificial Intelligence Conversational Agents in Health Care: Systematic Review. *J Med Internet Res*. 2020;22(10):e20346. doi:10.2196/20346
33. Laranjo L, Dunn AG, Tong HL, et al. Conversational agents in healthcare: a systematic review. *J Am Med Inform Assoc*. 2018;25(9):1248-1258.
34. Conversational agents in healthcare: a systematic review | Journal of the American Medical Informatics Association | Oxford Academic. Accessed June 28, 2023. https://academic-oup-com.ezproxy.neu.edu/jamia/article/25/9/1248/5052181
35. Rutjes H, Willemsen MC, IJsselsteijn WA. Beyond Behavior: The Coach's Perspective on Technology in Health Coaching. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Association for Computing Machinery; 2019:1-14. doi:10.1145/3290605.3300900
36. Jones RL, Wallace M. Another bad day at the training ground: Coping with ambiguity in the coaching context. *Sport Educ Soc*. 2005;10(1):119-134. doi:10.1080/1357332052000308792
37. Santini S, Fabbietti P, Galassi F, et al. The Impact of Digital Coaching Intervention for Improving Healthy Ageing Dimensions among Older Adults during Their Transition from Work to Retirement. *Int J Environ Res Public Health*. 2023;20(5):4034. doi:10.3390/ijerph20054034
38. Diego Arguello. Detecting and modifying daily inactivity among adults over 60 years using an integrated two-way communication-based near-real-time sensing system: A randomized clinical trial. Conference Presentation presented at: 8TH INTERNATIONAL CONFERENCE ON AMBULATORY MONITORING OF PHYSICAL ACTIVITY AND MOVEMENT; June 22, 2022; Keystone, CO, USA. https://ismpb.org/wp-content/uploads/2022/07/ICAMPAM-Program-2022_post-updates-1-1.pdf
39. Chase H. LangChain. Published online October 2022. Accessed July 6, 2023. https://github.com/hwchase17/langchain
40. Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Published online August 27, 2019. doi:10.48550/arXiv.1908.10084
41. Chroma: the AI-native open-source embedding database. Accessed July 10, 2023. https://www.trychroma.com
42. Clifford, Dawn, Curtis, Laura. *Motivational Interviewing in Nutrition and Fitness*. Guliford Press; 2016. Accessed December 6, 2023. https://www.guilford.com/books/Motivational-Interviewing-in-Nutrition-and-Fitness/Clifford-Curtis/9781462524181

43. Gururangan S, Marasović A, Swayamdipta S, et al. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. Published online May 5, 2020. doi:10.48550/arXiv.2004.10964
44. Wang L, Mujib MI, Williams J, Demiris G, Huh-Yoo J. An Evaluation of Generative Pre-Training Model-based Therapy Chatbot for Caregivers. Published online July 27, 2021. doi:10.48550/arXiv.2107.13115
45. Rashkin H, Smith EM, Li M, Boureau YL. Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset. Published online August 28, 2019. doi:10.48550/arXiv.1811.00207
46. Raffel C, Shazeer N, Roberts A, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Published online July 28, 2020. doi:10.48550/arXiv.1910.10683
47. Wei J, Wang X, Schuurmans D, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Published online January 10, 2023. Accessed July 10, 2023. http://arxiv.org/abs/2201.11903
48. Arakawa R, Yakura H. Coaching Copilot: Blended Form of an LLM-Powered Chatbot and a Human Coach to Effectively Support Self-Reflection for Leadership Growth. In: *Proceedings of the 6th ACM Conference on Conversational User Interfaces*. CUI '24. Association for Computing Machinery; 2024:1-14. doi:10.1145/3640794.3665549
49. Abd-Alrazaq A, Safi Z, Alajlani M, Warren J, Househ M, Denecke K. Technical Metrics Used to Evaluate Health Care Chatbots: Scoping Review. *J Med Internet Res*. 2020;22(6):e18301. doi:10.2196/18301
50. Denecke K, Abd-Alrazaq A, Househ M, Warren J. Evaluation Metrics for Health Chatbots: A Delphi Study. *Methods Inf Med*. 2021;60(5-06):171-179. doi:10.1055/s-0041-1736664
51. Hua Y, Liu F, Yang K, et al. Large Language Models in Mental Health Care: a Scoping Review. Published online January 1, 2024. doi:10.48550/arXiv.2401.02984
52. Nazi ZA, Peng W. Large Language Models in Healthcare and Medical Domain: A Review. *Informatics*. 2024;11(3):57. doi:10.3390/informatics11030057
53. Moyers TB, Martin T, Manuel JK, Hendrickson SML, Miller WR. Assessing competence in the use of motivational interviewing. *J Subst Abuse Treat*. 2005;28(1):19-26. doi:10.1016/j.jsat.2004.11.001
54. Moyers TB, Martin T, Manuel J, Miller WR. Motivational interviewing treatment integrity (MITI) coding system. *Univ N M Cent Alcohol*. Published online 2005.
55. Imel ZE, Baldwin SA, Baer JS, et al. Evaluating therapist adherence in motivational interviewing by comparing performance with standardized and real patients. *J Consult Clin Psychol*. 2014;82(3):472-481. doi:10.1037/a0036158
56. Green P, MacLeod CJ. SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods Ecol Evol*. 2016;7(4):493-498. doi:10.1111/2041-210X.12504

Facilities and Resources

Scientific Environment

Founded in 1898, Northeastern University is committed to providing the necessary resources for the research to their research active faculty. Northeastern University was ranked #53 out of all National Universities, and #35 for Computer Science in the U.S. News and World Report 2023 rankings. Northeastern University has an enrollment of more than 30,000 undergraduate and graduate students and approximately 4,300 faculty and staff. There are eight undergraduate colleges, eight graduate and professional schools and two undergraduate evening divisions. It is recognized as one of the top universities in the nation for practice-oriented education, which is largely based on the co-operative education program that is required for all undergraduate university students. Northeastern achieved R-1 research activity status in 2015, with $140.7 million in external research funding. Since then, research funding has increased almost 2-fold, with $230.7 million in external awards in 2022.

The facilities and resources available to Ethan at Northeastern University include what is required to successfully complete the proposed project. Ethan is currently enrolled in the PhD program in Personal Health Informatics, a joint offering from the Bouvé College of Health Sciences and the Khoury College of Computer Sciences. The Health Sciences Department, within the Bouvé College of Health Sciences, includes extramurally funded diverse faculty with expertise in the fields of exercise science, population health, health disparities, nutritional epidemiology, social epidemiology, medical sociology, public policy and mental health. As of 2023, the Khoury College of Computer Sciences is home to 190 faculty in a variety of computational fields. Many of the Personal Health Informatics faculty have joint appointment in both colleges, fostering a trans-disciplinary intellectual environment that incorporates academics, research, and practice.

Facilities

Laboratory and Office

The Human Performance and Exercise Science Laboratory located within the Behrakis Health Science Center, is equipped with 2 offices for research faculty and graduate research assistants. This laboratory occupies 2000 sq. ft. of space. The laboratory is designed to easily create private enclosures to perform various testing procedures on research study participants. The lab has various equipment to measure aerobic and strength capacity and various health-related outcome measures including body composition, aerobic and musculoskeletal fitness, and blood biomarker analyses. Currently the lab stores wearable computing devices like smartwatches, as well as housing computers and peripherals to deploy wearable devices with participants, download accelerometry data, and visualize activity data. More recently lab has served as a collaborative hub for multiple projects, hosting participant visits and project meetings. Ethan has been heavily involved in conducting various research projects in the lab with Dr. John, Dr. Intille, and other graduate students.

Computing

The Khoury College of Computer Sciences at Northeastern University provides exclusive computing resources for both faculty research and teaching purposes. These facilities are constantly being upgraded to keep pace with developments in the computer industry. Khoury's infrastructure is managed by the Khoury Systems team that is dedicated to computing support for the college and consists of five experienced Linux, Windows, Network, and Virtualization administrators with a dedicated Service Desk. Khoury Systems provides enterprise-grade server hosting in comprising of network, power, and cooling services. Researchers may utilize these capabilities by staging personal or research group computing systems in the Khoury Systems data center. Khoury Systems will also provide guidance in vendor selection, defining equipment specifications, and assisting in the purchasing process for all computing hardware requirements. Additionally, centrally managed applications such as GitHub Enterprise, WordPress Multisite, and Confluence Wiki are provided as tools for both research and teaching purposes.

Research Computing

Northeastern ITS Research Computing provides high-end research computing resources to all Northeastern University researchers and manages Northeastern's involvement in the Massachusetts Green High Performance Computing Center facility. These resources include centralized high-performance computing (HPC) clusters and storage, visualization, software, high-level technical and scientific user support, education, and training. Researchers can leverage these for both HPC and non-HPC research computing use, development, and funding requirements. As of 2023, the shared Discovery cluster provides secure access to over 20,000 compute cores, a mix of 10Gbps Ethernet and 56 Gbps FDR Infiniband network backplane and over 3 PiB of high-speed parallel file system storage. In addition, over 200 GPUs and a 50TB Hadoop experimental environment for big data research are available. **The computational resources available**

through the Discovery Cluster, specifically the cluster's graphical computing power and storage capability, are integral to the success of this project. Access to the Discovery Cluster's performance GPUs will accelerate the prototyping and development of the machine learning techniques used in this project. Project storage and other services are available to faculty members, with flexible options for data intensive projects. A dedicated team of research computing staff (and several graduate students) manages the environment and supports researchers in their use of high-performance computing for research and discovery.

Other resources:

The Bouvé College Office of Research (BCORe) assists faculty with the development of research proposals and protocols. The office also provides licensing for Qualtrics, an online secure data acquisition and management system that is available to all faculty, researchers, and students at the Bouvé College. Forms are easily created and can include data checks and skip, display, and branch logic. The program meets general HIPPA regulations and is complaint with standard security protocols. The college supports licensing costs. College faculty also has access to the REDCap (Research Electronic Data Capture) survey and data management system. REDCap is installed behind a secure and monitored firewalled server at Northeastern and can only be accessed via Northeastern's on-campus network or via remote virtual private network (VPN).

The Bouvé College Office of Research also supports the Biostatics Service Center (BSC) where faculty can get assistance with the development and review of analytic plans and statistical and database support for research projects. Analytic support for projects is performed on a fee-for-service basis with charges and payments coordinated through BCORe. Hourly rates are based on the level of expertise required. Support services are provided by professionally trained biostatisticians and by Northeastern University graduate students with demonstrated expertise in biostatistical applications. Work on all projects is reviewed and monitored by the senior biostatistical staff at the College.

Library

The Snell Library houses close to one million volumes, 2.3 million microforms, 23,000 audio, video, and software titles, and current subscriptions to more than 8,200 serials and newspapers. The Digital Media Commons features an innovative media lab where faculty can utilize a range of technologies and services including support for course design and development, digital scholarship, media creation and integration, digital archival tools, text encoding, geographic information systems tools, software tutorials, and project- management assistance. The facility's data-analysis capabilities and computer-aided design tools, fully supported by librarians and IS technical experts, provide users with the ability to work across disciplinary boundaries to explore innovative solutions to real-world problems. The library also provides access to the Current Contents Connect (Thomson Reuters/Web of Knowledge) that allows faculty to browse tables of contents of scholarly publications within a single web site and to subscribe to tables of contents of journals to receive an update when new issues are published.

Animal: Not applicable.

Clinical: Not applicable.

Equipment

As the goals of this project are to create and evaluate large language models for health applications, equipment needed to conduct this research is mostly limited to computing resources.

The Khoury College of Computer Sciences at Northeastern University provides exclusive computing resources geared towards both faculty research and teaching purposes. These facilities, subject to regular upgrades, encompass enterprise-grade server hosting services managed by the Khoury Systems team. This includes network, power, and cooling services necessary for hosting personal or research group computing systems. Researchers can also benefit from centrally managed applications, such as GitHub Enterprise, WordPress Multisite, and Confluence Wiki, offered as tools for both research and teaching purposes.

In conjunction with the college's facilities, Northeastern ITS Research Computing offers high-end research computing resources. The shared Discovery cluster, available as of 2023, has over 20,000 compute cores, a mix of 10Gbps Ethernet and 56 Gbps FDR Infiniband network backplane, and over 3 PiB of high-speed parallel file system storage. Crucially, the cluster is equipped with a select of over 200 NVIDIA K80, P100, V100, and T4 GPUs, **providing substantial graphical computing power crucial for the prototyping and development of machine learning techniques essential to my proposed project.** Storage services are also available, accommodating the substantial data generated by the research projects with flexible options for data-intensive projects.

Aside from physical infrastructure, the Discovery Cluster provides hundreds of software systems and packages. Most notable for this project, Discovery Cluster provides environmental and versioning management systems (Git and Conda), research computing languages (R and Python), SQL/NoSQL database systems, and parallelization packages (OpenMPI and CUDA). Hosting these applications is the Discovery Cluster's operating system, Linux-based CentOS. CentOS Linux is ideal for research computing due to its reliability and speed. Discovery Cluster also provides training and on-demand support for all users. Consulting is also available.

This comprehensive suite of computing resources, spanning wearable devices, laboratory computers, HPC access at the Khoury College of Computer Sciences, and the extensive Research Computing infrastructure, will play a pivotal role in data collection, analysis, and the development of advanced machine learning techniques central to the proposed research. I will additionally benefit from the support provided by a dedicated team of research computing staff, ensuring seamless integration and efficient use of these resources.

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors.
Follow this format for each person. **DO NOT EXCEED FIVE PAGES.**

NAME: Rogers, Ethan B.

eRA COMMONS USER NAME (credential, e.g., agency login): [redacted]

POSITION TITLE: Graduate Student Research Assistant

EDUCATION/TRAINING

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	Start Date MM/YYYY	Completion Date MM/YYYY	FIELD OF STUDY
University of Vermont, Burlington, VT	B.S.	08/2013	05/2017	Microbiology
Northeastern University, Boston, MA	Ph.D.	09/2021	05/2027	Personal Health Informatics

A. Personal Statement

In high school I took a Python programming class that opened the door to internships at the Center for Addiction Management (CAM). CAM needed data collection tools that could be run in MRI machines to provoke specific states of mind for their research on decision making. I had the knowledge to translate their needs into code, and I spent two summers building these tools and exploring the basics of scientific research. Propelled by this early exposure to scientific research at CAM and a general interest in health, I followed my curiosity to the University of Vermont (UVM) where I earned a B.S. in microbiology, worked with students as a teaching assistant, explored *C. difficile* sporulation in Dr. Aimee Shen's lab, and worked as a data intern for the Vermont Department of Health recording cases of food poisoning and summarizing seasonal trends among age groups. These opportunities gave me a solid foundation in biology and allowed me to further explore computing in health applications. After my undergraduate education, I worked at the Vermont Center for Behavior and Health as a research assistant gaining experience working with historically marginalized populations on large clinical trials relating to smoking behavior, before transitioning to the Vermont Child Health Improvement Program (VCHIP). Both principal investigators and heads of their own teams at VCHIP, Dr. Valerie Harder and Dr. Charles Mercier saw potential in my ability to apply my programming skills to automate and expedite data reporting and analyses. My work at VCHIP involved building data collection and reporting systems and producing scientific reports detailing the achievements and shortfalls of perinatal healthcare in Vermont. I was able to apply my problem solving and computing skillset towards a common goal within a group of diverse researchers, which laid a strong foundation for my ongoing doctoral studies. Currently I am a graduate research assistant, in my 4th year of Northeastern University's Personal Health Informatics doctoral program, advised by Dr. Dinesh John. Although I am no longer harvesting *Clostridium* spores or asking Vermonters about their smoking behavior, my experience and education developed in me a unique perspective on how computer science can be leveraged to solve health problems. In my work I use my background in biology, strong interest in human health, years of experience using code to solve problems, and many hours of working with participants to further my research goals: creating research tools which enable greater exploration of human health. Building systems which measure and promote health and wellness is an expression of my creativity and values. For example, during my first two years in my program, I used my technical background and experience building data analysis systems to develop a data processing pipeline to classify activity from 12 person-years of raw accelerometer using activity classification algorithms developed by Dr. Stephen Intille's (co-sponsor) lab. With these raw data transformed into meaningful predictions of activity, our lab found that augmenting exercise programs with socially engaging, and context-aware health coaching reduced sitting and increased standing behavior. This fellowship opportunity will allow me the opportunity to work with experts in the field of AI (Dr. Dakuo Wang) and behavioral interventions (Dr. Stephen Intille, Dr. Rober Leeman), the freedom to explore the areas of research most relevant to my interests, the training and experience I need to continue creating and evaluating novel health systems like that which is detailed in the project, and springboard a career rooted in my values and interests.

B. Positions, Scientific Appointments and Honors

Positions and Scientific Appointments

2021 – Present	Graduate Research Assistant, Northeastern University
2023 – 2024	Graduate Teaching Assistant, Northeastern University
2018 – 2021	Research Assistant and Program Coordinator, Vermont Child Health Improving Program, University of Vermont Medical School
2017 – 2018	Research Assistant, Vermont Center for Behavior and Health, University of Vermont Medical School
2016 – 2017	Data Intern, Enteric Disease Surveillance, Vermont Department of Health
Summer 2016	Course Assistant, Biology of Parasitism, Marine Biological Laboratory

Honors

2016	Teaching Assistant of the Year, Department of Microbiology and Molecular Genetics, University of Vermont
2013-2017	Presidential Scholarship, University of Vermont

C. Contributions to Science

1. **High School Research:** For two summers in high school, I was a research assistant at Massachusetts General Hospital, in the Center for Addiction Medicine creating computer programs for data collection. I designed and built an online data collection tool for a study on the effect of *Cannabis* on delay discounting and coded data collection tools in Python. My largest project was an adaptation of the computer game Pong with variable game-play parameters to elicit different mental states for research on fMRI brain activity of cooperation and competition. I created journal templates in LaTeX for research publication and assisted with participant data collection in ongoing research. The work I did these summers was my first introduction to scientific research, bridging my interest in computing and biology.
2. **Undergraduate:** As an undergraduate I was a teaching assistant for 3 semesters, working with professors to teach and administrate introductory microbiology and genetics wet lab courses. I was hired as a course assistant for Biology of Parasitism at the Marine Biological Institute during the summer of 2016, where I managed the course's wet lab. I did one semester of undergraduate research in a lab run by Dr. Aimee Shen studying them mechanisms of *C. difficile* sporulation but decided to follow my interest in public health and data by spending a year as the data intern for the Enteric Disease Surveillance division at the Vermont Department of Health (VDH). At VDH I cataloged enteric disease case reports, maintained and updated the enteric disease surveillance database, and analyzed historical enteric disease burden across age groups.
3. **After Undergraduate:** After college I worked at the Vermont Center for Behavior and Health at UVM as a research assistant on an experimental psychiatry study investigating the effect of low nicotine cigarettes on historically marginalized populations. Aside from direct participant data collection, I was the data quality coordinator. I moved within UVM to the Vermont Child Health Improvement Program (VCHIP), where I was a research assistant under Drs. Valerie Harder and Charles Mercier. I provided data and administrative support for projects using insurance claims and electronic medical records to investigate the impact of opioid prescribing policy on opioid overdose (resulting co-authorship of a manuscript) and using Vermont's birth certificate database and VCHIP's relationships with perinatal clinicians to evaluate and improve perinatal care quality at community birthing centers around Vermont.
    1. Harder, V. S., Plante, T. B., Koh, I., **Rogers, E. B.**, Varni, S. E., Villanti, A. C., Brooklyn, J. R., & Fairfield, K. M. (2021). Influence of Opioid Prescription Policy on Overdoses and Related Adverse Effects in a Primary Care Population. *Journal of General Internal Medicine*, *36*(7), 2013–2020. https://doi.org/10.1007/s11606-021-06831-4
4. **Graduate:** I am currently a graduate research assistant at Northeastern University advised by Dr. Dinesh John. My most recent project applied machine learning algorithms to predict activity from accelerometer data collected during a pilot study investigating how activity-aware personalized health coaching affects unstructured physical activity in older adults. I have coauthored a methods manuscript[1] with my colleges and presented two posters at the *International Conference on Ambulatory Monitoring of Physical Activity and Movement (ICAMPAM)* in 2022[2] on this work. Additionally, I have explored the use of commercial

Large Language Models (LLMs) as a method to automate semantic tagging of messages without the need for large training datasets or advanced NLP expertise. Not only was I able to gain experience working with LLMs but I was also present this work as a poster at ICAMPAM 2024[3]. Aside from using the activity classifications to determine outcomes, I am in the process of publishing a paper on the system I designed track and process the 12 person-years of raw accelerometry data collected as part of my lab's research as well as a statistical methods paper on the use of mixed models in complex designs.

1. Arguello, D., **Rogers, E**., Denmark, G. H., Lena, J., Goodro, T., Anderson-Song, Q., Cloutier, G., Hillman, C. H., Kramer, A. F., Castaneda-Sceppa, C., & John, D. (2023). Companion: A Pilot Randomized Clinical Trial to Test an Integrated Two-Way Communication and Near-Real-Time Sensing System for Detecting and Modifying Daily Inactivity among Adults >60 Years—Design and Protocol. *Sensors*, *23*(4). https://doi.org/10.3390/s23042221
2. **Rogers, E.** *et al.*, "An open-source and automated data processing and reporting pipeline for continuous wearable data in adaptive interventions," presented at the *8th International Conference on Ambulatory Monitoring of Physical Activity And Movement*, Keystone, CO, USA*,* Jun. 2022.
3. **Rogers, E.**, Arguello, D., Anderson-Song, Q., & John, D., "Exploring the Feasibility of Commercial Large Language Models for Semantic Tagging of Health Coach Messages," presented at the *9th International Conference on Ambulatory Monitoring of Physical Activity And Movement,* Rennes, Brittany, France*,* Jun. 2024.

D. Scholastic Performance

YEAR	COURSE TITLE	GRADE
	UNIVERSITY OF VERMONT	
2013	Exploring Biology 1	[redacted]
2013	General Chemistry 1	
2013	Fundamentals of Calculus I	
2013	TAP: Health Care in America	
2014	Exploring Biology 2	
2014	General Chemistry 2	
2014	Fundamentals of Calculus II	
2014	General Psychology	
2014	D2: Prehistoric Archaeology	
2014	Organic Chemistry 1	
2014	Microbiology & Infectious Disease	
2014	Psychology Research Methods I	
2015	Organic Chemistry 2	
2015	D1: Race & Ethnicity Literature	
2015	Intro Recombinant DNA Tech	
2015	Molecular and Cell Biology	
2015	Basic Statistical Methods	
2015	Biochemistry I	
2015	Teaching Assistant (Microbiology 101)	
2015	Introduction to Pharmacology	
2015	Effective Speaking	
2015	Medical Biostatistics & Epidemiology	
2016	Biochemistry of Human Disease	
2016	Immunology	
2016	Undergraduate Research (Microbiology)	
2016	Clinical Microbiology I	
2016	Toxicology	
2016	Genetics	
2016	Elementary Japanese I	

YEAR	COURSE TITLE	GRADE
2016	Prokaryotic Molecular Genetics	[redacted]
2016	Teaching Assistant (Microbiology 101)	
2016	Internship (Marine Biological Laboratory, Biology of Parasitism, Course Assistant)	
2016	Senior Seminar	
2017	Methods in Bioinformatics	
2017	Wilderness First Responder	
2017	Independent Study (Vermont Department of Health, Enteric Disease Surveillance, Data Intern)	
2017	Teaching Assistantship (Intro to Recombinant DNA Technology)	
2017	Emergency Medicine Research I	
2017	Advanced Topics in EMS Research	
	UNIVERSITY OF VERMONT (CONTINUING EDUCATION)	
2019	Climate Change Emergencies	
2020	Calculus I	
2020	Calculus II	
2020	Applied Linear Algebra	
2021	Advanced Bioinformatics	
	NORTHEASTERN UNIVERSITY	
2021	Theory and Methods in Human-Computer Interaction	[redacted]
2021	Theoretical Foundations in Personal Health Informatics	
2022	Programming Design Paradigm	
2022	Special Topics in Human Centered Computing: Human Sensing	
2022	Empirical Research Methods	
2022	Statistical Methods for Computer Science	
2023	Evaluating Health Technologies	
2023	Readings	
2023	Readings	
2023	Readings	
2023	Teaching Assistantship (Statistical Methods for Computer Science)	
2023	Database Management Systems	
2024	Teaching Assistantship (Computer Science Research Methods Seminar)	
2024	Natural Language Processing	
2024	Readings	
2024	Readings	

PHS Fellowship Supplemental Form

OMB Number: 0925-0001
Expiration Date: 01/31/2026

Introduction	
1. Introduction to Application (for Resubmission applications)	Introduction_to_Resubmission_RESUB_FINAL.pdf
Fellowship Applicant Section	
2. Applicant's Background and Goals for Fellowship Training*	Background_Goals_Fellowship_Training_RESUB_FINAL.pdf
Research Training Plan Section	
3. Specific Aims*	Specific_Aims.pdf
4. Research Strategy*	Research_Strategy_RESUB_FINAL.pdf
5. Respective Contributions*	Respective_Contributions_RESUB_FINAL.pdf
6. Selection of Sponsor and Institution*	Selection_of_Sponsor_and_Institution_V1_DJ.pdf
7. Progress Report Publication List (for Renewal applications)	
8. Training in the Responsible Conduct of Research*	Responsible_Conduct_of_Research_RESUB_FINAL.pdf
Sponsor(s), Collaborator(s) and Consultant(s) Section	
9. Sponsor and Co-Sponsor Statements	SponsorCosponsor_Resub_DW_SSI_RESUB_FINAL.pdf
10. Letters of Support from Collaborators, Contributors and Consultants	Leeman_Rogers_LoS_RESUB_FINAL.pdf
Institutional Environment and Commitment to Training Section	
11. Description of Institutional Environment and Commitment to Training	Institute_Environ_CommitTraining_RESUB_FINAL.pdf
12. Description of Candidate's Contribution to Program Goals	
Other Research Training Plan Section	

Vertebrate Animals

The following item is taken from the Research & Related Other Project Information form and repeated here for your reference. Any change to this item must be made on the Research & Related Other Project Information form.

Are Vertebrate Animals Used? ☐ Yes ☑ No

13. Are vertebrate animals euthanized?
   If "Yes" to euthanasia
   Is method consistent with American Veterinary Medical Association (AVMA) guidelines?
   If "No" to AVMA guidelines, describe method and provide scientific justification

14. Vertebrate Animals

PHS Fellowship Supplemental Form

Other Research Training Plan Information

15. Select Agent Research

16. Resource Sharing Plan

17. Other Plan(s)

18. Authentication of Key Biological and/or Chemical Resources

Additional Information Section

19. Human Embryonic Stem Cells

Does the proposed project involve human embryonic stem cells?* ☐ Yes ☑ No

If the proposed project involves human embryonic stem cells, list below the registration number of the specific cell line(s), using the registry information provided within the agency instructions. Or, if a specific stem cell line cannot be referenced at this time, please check the box indicating that one from the registry will be used:

Specific stem cell line cannot be referenced at this time. One from the registry will be used.

Cell Line(s):

20. Alternate Phone Number:

21. Degree Sought During Proposed Award:

Degree:	If "other", indicate degree type:	Expected Completion Date (MM/YYYY):

22. Field of Training for Current Proposal*: 419 Computer & Information Science, Other

23. Current or Prior Kirschstein-NRSA Support?* ☐ Yes ☑ No

If yes, identify current and prior Kirschstein-NRSA support below:

Level*	Type*	Start Date (if known)	End Date (if known)	Grant Number (if known)

24. Applications for Concurrent Support?* ☐ Yes ☑ No

If yes, describe in an attached file:

25. Citizenship*

U.S. Citizen U.S. Citizen or Non-Citizen National? ☑ Yes ☐ No

Non-U.S. Citizen ☐ With a Permanent U.S. Resident Visa

☐ With a Temporary U.S. Visa

If you are a non-U.S. citizen with a temporary visa applying for an award that requires permanent residency status, and expect to be granted a permanent resident visa by the start date of the award, check here: ☐

26. ☐ Change of Sponsoring Institution

Name of Former Institution:*

Tracking Number: GRANT14229905

Opportunity Number: PA-23-272 . Received Date: 2024-08-08T13:58:38.000-04:00

PHS Fellowship Supplemental Form

Budget Section

All Fellowship Applicants:

27. Tuition and Fees*:

None Requested ✔ Funds Requested

Year 1

Year 2

Year 3

Year 4

Year 5

Year 6 (when applicable)

Total Funds Requested:

[redacted]

28. Childcare Costs*:

✔ None Requested Funds Requested

Year 1

Year 2

Year 3

Year 4

Year 5

Year 6 (when applicable)

Total Funds Requested: $0.00

Senior Fellowship Applicants Only:

29. Present Institutional Base Salary: Amount Academic Period Number of Months

30. Stipends/Salary During First Year of Proposed Fellowship:

a. Federal Stipend Requested: Amount Number of Months

b. Supplementation from Other Sources: Amount Number of Months

Type (e.g.,sabbatical leave, salary)

Source

Appendix

31. Appendix

1. Introduction. This application is a resubmission of F31HL176319, "*Building and evaluating scalable health coaching system with large language models to promote physical activity.*" The reviewers' feedback has been addressed in this revised application, and key points summarized thematically. I deeply appreciate the consideration and effort of the reviewers and have improved this submission with their feedback.

1.1 Areas of Improvement as a F31 Trainee.

As noted by each reviewer, at the time of the submission I did not have preexisting experience with the specific methods of this work - natural language processing (NLP), machine learning/AI, or behavior change – nor an undergraduate background in computer science (CS). Although this is mitigated by my experience teaching myself to program (*"This weakness was mitigated somewhat by his history of successful self-teaching"*) the reviewers noted formal training could *"amplify the candidate's practical skills."* **Anticipating this feedback, I have taken steps to gain experience and coursework in these areas since my I initial submission.**

- To better understand the methods of NLP and AI, I completed CS6210: Fundamentals of NLP (Graduate Level).
- To demonstrate my working knowledge of Large Language Models (LLMs), I presented a poster on the application of commercial Large Language Models for automated tagging of health coaching messages at the *9th International Conference on Ambulatory Monitoring of Physical Activity and Movement*. This work was peer reviewed before acceptance to the conference.
- To gain experience with motivational interviewing (MI) – the core theoretical approach of this work – I completed a Certificate of Intensive Training in Motivational Interviewing with UMass Chan Medical School. During this four-month program I attended lecture sessions, studied the theory and approach of MI, practiced MI skills with simulated patient encounters, and received personal feedback and coaching from MI interventionists.

1.2. Addition of Dr. Robert Leeman as an Advisor on Theories of Behavior Change

It was noted that my Sponsorship team lacked expertise on theories of behavior change. Dr. Robert Leeman, the Chair of the Department of Public Health and Health Sciences at Bouvé College of Health Sciences, has agreed to provide guidance on appropriately integrating theories of behavior change in the proposed work. Dr. Leeman has extensive experience designing and evaluating digital health interventions based in theories of behavior change and has pledged to meet monthly to discuss the application of these theories to my research.

1.2. Generalizability of Research and Efficacy of the Training Data

- *"It was not clear if the training [fine-tuning] data were representative of the larger population or how that would impact model refinement."* These data were collected from an intervention to improve physical activity (PA) of older adults. **In this resubmission, I refined the population to older adults**, as 1) with greater chronic illness, older adults have the greatest need of interventions to increase PA, and 2) the messages used to train the models in this intervention are already tailored to the needs and communication styles of older adults. I feel this will increase the generalizability of the work and maximize the utility of the AI system.
- *"[Regarding the use of health coach messages as the standard to emulate] What makes your ground truth ground truth?"* I have added evidence showing the intervention that generated the health coaching messages used for evaluation of the system increased light PA by around an hour per day.

1.3 Strength of the Evaluation. The evaluation lacked statistical detail and alignment with the goals of system – provide high quality responses with minimal hallucination.

- *"BERTscores were going to be compared across different models, but no mention was made referring to inferential statistics to be used to guide these comparisons."* I have outlined the statistical methods (ANOVA/mixed models), sample size justifications, and experimental design to compare model performance.
- *"What is the metric you are hoping to examine? How will you evaluate if [the system] avoids hallucination?"* The human evaluation of the system is now redesigned to harness the judgement of health coaches. Instead of just qualitative usability interviews, human evaluation now includes rating system responses for humanness, adherence to MI principles, and degree of hallucination by health coaches. This way I can better quantify the system's performance with metrics based on the perception of the users of the system – health coaches.

1.4 Other Improvements.

- *"Insuring data would not reveal characteristics of the participant was buried in the exemption statements. Additional detail needed around data protections and storage."* I have included a Data Management and Sharing plan to alleviate related concerns raised during review.
- *"No mention of 8 hours per year Responsible Conduct of Research training requirement or future actions*." I have revised the Responsible Conduct of Research training plan to include specific activities to meet the training requirements as well as the institutional support in place to ensure I meet this requirement.
- All three reviewers found the requested three years of support to be excessive for the proposed work. I have altered the funding period to two years. To better align with the amount of work in this proposal and expected graduation timeline.

Background and Goals for Fellowship Training

Doctoral Dissertation and Research Experience

Highschool and Undergraduate Research

My research career began late in high school when I spent two summers working under Dr. Jodi Gilman at the Center for Addiction Medicine at the Massachusetts General Hospital. Working with other research assistants and principal investigators, I built software solutions by incorporating feedback on the needs of our research team from both a scientific and UX/UI standpoint which made the data collection process more accurate and efficient. For example, I built data collection tools/packages designed to elicit states of competition or cooperation to measure the resulting brain activation in participants with and without substance use disorders.

Although I enjoyed the role of a technology specialist on the research team, I went on to earn a Bachelor of Science in microbiology at the University of Vermont (UVM). While at UVM, I worked as a teaching assistant for three semesters, conducted research with Dr. Aimee Shen on the mutations to the cleavage site of *Clostridium difficile*'s primary germination lytic transglycosylase, and worked to record and analyze case reports of enteric disease at the Vermont Department of Health. In the latter part of my time at UVM, I worked under senior epidemiologist Bradley Tompkins at the Vermont Department of Health. In this role, I entered epidemiological reports from laboratory-confirmed cases of the four most common enteric diseases in the state. This opportunity, along with my previous experience in academic labs, cemented my interest in research and helped me to secure my first job after graduation. However, my experiences while obtaining my bachelor of science were instrumental in charting my future career path as, I realized that, though I am interested in the material of microbiology and health, I prefer computation and programming and interacting with human subjects and behavioral research over wet lab work.

Early Career Experience Working on Clinical Trials

After graduation in 2017, I was hired as a research assistant for a large clinical trial at the Vermont Center for Behavior and Health (VCBH). Under the direction of Dr. Stephen Higgins, this clinical trial studied the effects of low-nicotine cigarettes on the smoking habits of historically marginalized populations, particularly women, people with opioid use disorder, and people with mental health conditions. Though it has the authority to do so, the FDA had never regulated the amount of nicotine in cigarettes, primarily because the effects, especially on people more susceptible to the disruption caused by nicotine withdrawal, had not been fully examined. My responsibilities at VCBH included completing weekly check-ins with participants who volunteered to use low-nicotine cigarettes, at which we would discuss and record their cigarette use, substance use, health, and their experiences in the study. Working with historically marginalized populations is challenging and taught me the importance of staying level-headed and maintaining the ability to communicate effectively and with empathy for the participant. Working closely with participants and completing my other responsibilities at VCBH enabled me to develop a broad collection of research experiences: designing tools for research, working in a wet lab, analyzing previously collected data, and now working directly with study participants.

Vermont Child Health Improvement Program

After working for a year at VCBH, I transitioned into a role as a research assistant at the Vermont Child Health Improvement Program (VCHIP), a community-focused organization part of UVM's Larner College of Medicine tasked with bringing together healthcare providers, state health officials, and community leaders to improve healthcare for Vermont's children. Working in my role at VCHIP was the first time that I was able to fully practice my unique combination of skills and background knowledge to further the work of my team and our research.

I worked on several projects at VCHIP, of which, two of the most consequential projects were the Vermont Regional Perinatal Health Project (VRPHP), under Dr. Charles Mercier, and a study on the effects of legislation meant to curb opioid addiction, under Dr. Valerie Harder. Both projects allowed me to draw upon and expand my knowledge of computing and health. For example, I collaborated with state partners and the VRPHP team to combine the Vermont birth certificate registry with VRPHP's existing hospital statistics reporting system, expanding the scope of the project's perinatal health quality improvement efforts. Using the birth certificate registry, I released the first-ever statewide homebirth report, which compared the perinatal outcomes of home deliveries with those of hospital births. While working on the opioids study, I learned that it was possible to probe Vermont's all-payer claims database and the UVM Medical Center's electronic medical records to answer questions about opioid overdose, emergency room visits, and complications related to substance use disorder. Using these data, my team and I conducted an analysis, which revealed that although overdose among primary care patients did not decrease after the implementation of opioid prescribing limitations, adverse effects from opioids fell 78% from pre-policy levels. Our paper detailing this finding, my first publication, was published in the Journal of General Internal Medicine[1].

During my tenure at VCHIP, I developed other important skills, such as how to present information to a new audience and how to work closely with clinicians and other professionals who were unfamiliar with statistics and computing. I also gained experience with the procedural aspects of scientific communication: narrowing the scope of one's analysis, defining a research question, documenting your analysis and conclusions, and preparing manuscripts for publishing. The freedom to explore potential solutions and the faith my team members had in my ability to achieve our shared objectives, empowered my success at VCHIP and laid the foundation for my transition to my current doctoral program at Northeastern University. I left VCHIP with specific knowledge of opioid addiction, population-level data collection systems like the birth certificate registry and insurance claims database, and the practice of healthcare quality assessment and improvement. My career has benefitted from the experience of working with a diverse set of partners in government, the healthcare system, and academic research.

Interest in Computing

I have integrated computing and programming into every research position I have held. While working at the Vermont Department of Health as an undergrad, I used SAS macros to examine seasonal fluctuation in *campylobacter* infections and to perform data cleaning and schema updates on the case reports database. At VCHIP, I built an entire reporting system and database to fully transition data collection at our partner hospitals from faxed paper forms to a secure online portal, accelerating my team's access to data and availability of automatic analyses– all while maintaining the strict standards of confidentiality needed when working with HIPAA protected PHI. This task involved collaboration with stakeholders at hospitals, representatives from the Vermont Department of Health, downstream users of our data, and other teams at VCHIP. I had to identify and push for the change to our reporting infrastructure, then design and build multiple systems; each for a different end user with different needs. Finally, I had to manage the transition, educate my colleagues, and when I left for graduate school hand off administration to the rest of my team.

In each of my positions, I was the only programmer on the team. As a result, I almost always took on these tasks alone with only my own. This meant that until my first semester of Northeastern University's Personal Health Informatics doctoral program, everything I knew about programming, designing databases and data collection tools, and creating analysis pipelines and reporting systems was self-taught. At Northeastern I have filled in the gaps in my computational knowledge, taking classes on object oriented programming, natural language processing, macheine learning for human sensing, and database design. I believe that my experiential background in programming, graduate coursework in computation, and undergraduate education in biological science give me a unique perspective on computing and health. Combined with my experience at multiple levels of clinical research, from working with participants to analyzing previously collected datasets, I believe I am uniquely suited to explore questions at the interface of health and computer science.

Companion Trial and Northeastern University

I began Northeastern University's doctoral program in Personal Health Informatics in 2021 with Dr. Dinesh John as my mentor. I started work on an ongoing clinical trial, the Companion trial, studying the effect of personalized health coaching on reducing sedentary behavior in older adults. My programming skills, along with my previous experience working with participants, meant that I could assist with the daily operations of the Companion trial while designing a system to process terabytes of accelerometer data. Preliminary results from the companion trial indicate personalized context-aware health coaching motivates participants to sit less in favor of light physical activity[2]. I presented a poster of my work at the International Conference on Ambulatory Monitoring of Physical Activity and Movement (ICAMPAM) twice – in 2022 on the creation of an activity classification and data processing pipeline to receive, store, and analyze raw accelerometry data[3], and most recently in 2024 detailing the feasibility of commercial large language models for automated semantic tagging of text messages from health coaches[4].

At Northeastern University, under Dr. John's mentorship and guidance from Dr. Wang and Dr. Intille, I have proposed this project as a continuation of the research explored with the Companion trial. My patient-facing experience conducting clinical research, my background in health and biology, and my demonstrated expertise in research computing have positioned me well to bridge the gap between technology and health and given me a strong knowledge base to successfully conduct the proposed research. This project will guide and enhance my scientific training while exposing me to new ideas and offering opportunities to share my work.

Training Goals and Objectives

Currently, I am on track to finish the required coursework, teaching obligations, and qualifying exams by the fall of 2024, at which point I will be a PhD candidate. My coursework so far has focused on core components of Personal Health informatics, mainly theories of behavior change, statistics, research methods, human-computer interaction, and computer science.

This award will allow me the focus and resources to further develop and actively practice my skills in applying computer science to health on a project I had a role in designing and an area of research in which I have great interest. To date, my research and course work have focused on identifying health problems and developing technological solutions that improve understanding of one's personal health, access to health information, and agency to seek treatment. During my time at Northeastern, I have used my programming skills and experience with clinical trials to implement a data analysis pipeline that provides rich, high-level summaries of participant activity enabling personalized, activity-aware health coaching. With this fellowship, the guidance of my sponsor team, and the research outlined in this proposal, I will be able to craft a research foundation based on novel technologies with the power to vastly increase access to behavioral health interventions. I will have training opportunities to learn more about behavioral interventions, experimental design, and language models; resources and time to disseminate my work; and a strong research foundation on which to build my career.

The proposed project will involve the use of a previously collected text-message dataset focusing on physical activity between health coaches and participants to fine-tune a large language model (LLM) within an LLM-augmenter framework, a conceptual framework that layers additional models and prompt engineering to constrain LLMs from producing non-factual information. Working with LLMs requires the use of high-performance computing (HPC) systems, such as Northeastern's Discovery cluster. Although I have previous experience working with HPCs, including as part of the construction of an analysis pipeline to run activity classification algorithms on raw accelerometer data on HPCs, machine learning applications like LLMs require new technical skillsets such as scheduling and managing large resource-intensive jobs, using machine learning frameworks for graphical processing units, and storing and evaluating fine-tuned machine learning models. Northeastern's Discovery cluster provides resources and training for these skillsets, as well as on-demand support for all users. Aside from technical skillsets, this project will provide opportunities to learn about conceptual approaches like model selection, fine-tuning, prompt testing, and evaluation. LLMs are very diverse, each with their own strengths and weaknesses. LLMs also produce qualitative outputs that are hard to easily measure or quantify and can be fine-tuned to focus and extend the capability of a model for a specific task. Selection of the correct model and the correct evaluation schema requires experience working with these models and evaluation metrics, as well as a detailed understanding of the intended use. I have limited previous experience with model selection and evaluation, prompt design, fine-tuning from my research on on the use of commercial LLMs for semantic tagging of health coach messages, presented in my abstract at ICAMPAM 2024. The proposed project will allow me to gain more experience selecting, fine-tuning, and testing LLMs for health applications, all under the supervision and expertise of Dr. Wang of the Northeastern Human-Centered AI Lab.

Aside from the technical components of this research, this research will involve meeting with stakeholders, designing evaluation schemes, performing human subjects research, and statisical analyses. With his experience and connections in the field of health behavior change research and his work designing clinical trials, Dr. John's expertise and guidance will be essential when performing these research activities. Dr. John will guide and advise on the recruitment of health coaches, the experimental design of the evaluations, and the statistical analysis of the collected data. Dr. John will also help troubleshoot issues that will inevitably arise when conducting human subject research.

Dr. John and Dr. Intille also have extensive experience with theories of behavior change – the building blocks of effective interventions. The use of behavior change theories is necessary when crafting effective interventions, and to successfully build a useful language model for health coaching we must integrate behavior change theories like Self-Determination Theory and Motivational Interviewing (MI). Self-Determination Theory is a psychological framework that emphasizes the importance of intrinsic motivation and the fulfillment of basic psychological needs in driving human behavior, while MI is a client-centered counseling approach that helps individuals explore and resolve ambivalence about behavior change by eliciting their own motivations and goals. These theories were used in the original research, and therefore their approach and implementations are carried out in the conversations I will use to train the health coaching model. As MI/SDT is a fundamental component of health coaching, I have completed a *Certificate of Intensive Training in*

Motivational Interviewing at the University of Massachusetts at Dartmouth. This semester-long progra involved cousework on theory and components of MI, lectures given by research faculty studying MI, practice sessions with trained mock-patients, and individual coaching sessions with MI practicioners. With this training and guidance from Dr. Intille and Dr. John, I will be better able to communicate with health coaches on the practice of MI-based health coaching and integrate the health coaches' feedback into the proposed research.

Beyond the skills needed to conduct this research are the experiences to share and communicate my findings. Under the direction of my advisor, Dr. John, I have previous experience writing scientific papers and presenting my work at poster sessions. In his work and his mentorship, Dr. John emphasizes the importance of strong scientific communication skills. With his guidance, my scientific writing has matured, and my scientific thinking has become more precise. We regularly meet where he offers both specific and abstract feedback on my writing as well as guidance with research planning and research activities. As an interdisciplinary program, personal health informatics students must possess complementary expertise in the fields of health and computer science. We must be experts in health and biology, human subjects research, and behavior change; all while being experienced programmers capable of building user-facing systems with iterative design and development. This is reflected in our coursework, diverse faculty in both the Bouvé College of Health Sciences as well as the Khoury College of Computer Sciences, and our three qualifying exams – one for health, one for computer science, and the requirement we publish our work in a peer-reviewed journal before candidacy. This project will allow me the opportunity to present my work to a diverse audience comprised of other interdisciplinary scientists. It will allow me to work closer with my sponsor team, Dr. Intille and Dr. Wang, as well as the opportunity to continue improving my scientific writing. Dr. John's guidance, the support from other faculty in my program, and the opportunities afforded through this grant will enable my future career goal of doing translational research that applies technical solutions to health problems. I plan to seek a post-doctoral position that will enable me to continue my research building and evaluating AI health coaching systems.

Activities Planned Under this Award

Year 1 (2025 Q2 – 2026 Q2) Northeastern University

Breakdown: Research 90%, Conferences/manuscript writing/Career Development: 10%

Research: Complete Specific Aim 1A using methods outlined in the research strategy. Build and deploy working prototype of Companion 2. Complete Specific Aim 1B – the backend and user interface for the health coach messaging system with candidate responses. Start metric-based evaluation of Companion2 responses (Specific Aim 2A).

Conferences/manuscript writing/Career Development: Propose and begin dissertation. Attend *International Conference on Ambulatory Monitoring of Physical Activity and Movement* 2026 with oral presentation.

Year 2 (2026 Q2 – 2027 Q2) Northeastern University

Breakdown: Research 70%, Conferences/manuscript writing and submission/Career Development: 30%

Research: Complete Specific Aim 2A and Specific Aim 2B, including all metrics-based and health coach evaluation as well as usability testing with health coaches.

Conferences/Symposia/Career Development: Complete and Defend dissertation. Submit paper and attend *Association of Computing Machinery Conference on User Modeling, Adaptation, and Personalization 2027*.

References

1. Harder VS, Plante TB, Koh I, et al. Influence of Opioid Prescription Policy on Overdoses and Related Adverse Effects in a Primary Care Population. *J Gen Intern Med*. 2021;36(7):2013-2020. doi:10.1007/s11606-021-06831-4

2. Diego Arguello. Detecting and modifying daily inactivity among adults over 60 years using an integrated two-way communication-based near-real-time sensing system: A randomized clinical trial. Conference Presentation presented at: 8TH INTERNATIONAL CONFERENCE ON AMBULATORY MONITORING OF PHYSICAL ACTIVITY AND MOVEMENT; June 22, 2022; Keystone, CO, USA. https://ismpb.org/wp-content/uploads/2022/07/ICAMPAM-Program-2022_post-updates-1-1.pdf

3. Ethan Rogers, Diego Arguello, Grant Denmark, et al. An open-source and automated data processing and reporting pipeline for continuous wearable data in adaptive interventions. In: International Society for the Measurement of Physical Behaviour; 2022.

4. Ethan Rogers, Diego Arguello, Quinn Anderson-Song, Dinesh John. Exploring the Feasibility of Commercial Large Language Models for Semantic Tagging of Health Coach Messages. In: International Society for the Measurement of Physical Behaviour; 2024.

SPECIFIC AIMS

Older Americans are both less likely to meet physical activity recommendations[1] and more likely to have chronic illness[2] than younger Americans. Because lack of physical activity (PA) contributes to preventable chronic illness, older adults need tailored interventions that target their specific barriers to PA[1,3]. Personalized physical activity health coaching (PAHC) can increase daily PA in older adults[3–5] thereby reducing the risk of Alzheimer's Disease and other neurodegenerative illness[6] and contributing to healthy aging[7]. PAHC leverages the personal relationship between the coach and the client to establish goals, increase accountability, troubleshoot barriers, and celebrate achievements to motivate behavior change. However, establishing this relationship and being responsive takes time and effort, limiting the scalability of health coaching in both the number of clients and frequency of contact. Human-in-the-loop artificial intelligence (AI) PAHC, where an AI health coach generates responses that are reviewed, edited, and sent by human health coaches can expand the reach of human PAHC to more clients by partially automating the generation of meaningful responses. Early health coach chatbots had limited ability to respond to unexpected responses or parse personal information from a client's responses, leading to a rigid and unengaging dialog where personal details were lost or underutilized. This limited the chatbot's ability to replicate the coach-client relationship, thus reducing its efficacy at promoting behavior change[8]. As AI has matured, new technologies better suited to natural dialog have emerged. Large Language Models (LLMs) like ChatGPT and LLaMA have demonstrated emergent capabilities in question answering, sentiment analysis, and text completion. These models produce sophisticated, meaningful, responses to open-ended prompts and under the right conditions, appear convincingly human, making them excellent candidates for AI PAHC. In this work I will address the scientific gaps preventing adoption of LLMs for PAHC: lack of message sets to fine-tune LLMs for PAHC and reducing the propensity of LLMs to hallucinate. Although LLMs are highly suited to chat applications that require the ability to respond dynamically to any input, acquiring a PAHC dataset for model fine-tuning to adapt LLMs to specific tasks is a barrier to the adoption of LLMs for health coaching applications[8,9]. Through previous research[10], my lab has amassed a 20k PAHC message dataset between health coaches and 23 older clients (mean age of 72) suitable for model fine-tuning. Another concerning limitation for health applications is the propensity of LLMs to 'hallucinate' or generate non-factual statements. To mitigate hallucination, I will employ the LLM-Augmenter framework to query a curated document library for up-to-date knowledge and focus the scope of the model's response[11]. The overall purpose of this work is to build and evaluate a hybrid AI PAHC system that can replicate the responses of human PA health coaches, while addressing the deficiencies of LLMs with a PAHC fine-tuning dataset and LLM-Augmenter to reduce hallucination. I *hypothesize* that combining an LLM fine-tuned with our PAHC message set with an LLM-Augmenter will yield a scalable hybrid system that replicates the quality and personalization of manual PAHC. This hybrid system has the potential to increase the scalability and reach of a PA coach, while maintaining a high quality of coaching, thus greater likelihood of behavior change. We will achieve this with the following specific aims.

Aim 1: Construct a hybrid, human-in-the loop system for AI PAHC.

First, for **Aim 1A** I will build a system capable of replicating human PAHC responses focused on older adults. I will fine-tune a LLaMA 2 model[12] using our existing PAHC message set described earlier. To integrate external knowledge and reduce hallucination I will use the LLM-Augmenter framework[11,13]. This framework layers additional LLMs which determine 1) what external information is needed to respond to user input, 2) retrieve that information from a curated database and, 3) provide that context with an appropriate prompt to the fine-tuned (on the 20k PAHC message set) model for its final response to the user. This chained architecture has been shown to mitigate hallucination in medical chatbots[13]. Second, for **Aim 1B** I will construct a system backend and user interface that enables health coaches to review, edit, and send AI-generated responses for human-in-the-loop PAHC messaging.

Aim 2: Evaluate the system's ability to replicate manual PAHC responses with metrics, human evaluation, and usability interviews (N=5).

Evaluation consists of two methodologies, metrics-based evaluation as **Aim 2A** and human evaluation as **Aim 2B**. For metrics-based evaluation I will use BERTscore[14] to test the similarity of the human and AI health coaches' responses. BERTscore can capture the intent and meaning similarity between two text inputs, allowing a comparison of the semantic similarity between a health coach message from our message set and the AI PAHC system's response when given the same conversation history. I will recruit five PA-focused health coaches for human evaluation of the system. Health coaches will rate the humanness, behavior change methods adherence, and degree of hallucination present in the AI PAHC system's responses. Coaches will have the opportunity to test the interface, and I will conduct qualitative interviews on the usability of the human-in-the-loop system interface to inform future design changes.

RESEARCH STRATEGY

1. Significance.

Daily physical activity (PA) is a method of prevention and treatment for many non-communicable chronic diseases[15–18]. PA can reduce the risk of cancer[19,20], cardiovascular disease[21], stroke[22], and diabetes[23]. There is clear evidence that daily moderate-to-vigorous PA decreases the risk of chronic disease, but any intervention that decreases sedentary behavior benefits health outcomes[24,25]. Specifically relevant to older adults is the relationship between PA and neurodegenerative illness. Lack of PA has been identified as a risk factor for all forms of dementia[6,26] regardless of susceptibility. Conversely, interventions that promote increased PA have proven effective in slowing cognitive decline in individuals with Alzheimer's Disease or other neurodegenerative diseases[27]. Aside from neurodegenerative disease, increased PA is associated with greater life satisfaction while aging[7]. Unfortunately, less than 15% of American older adults meet the Physical Activity Guidelines[1], thus, interventions that aim to increase PA among older adults are required to prevent and slow the progression of chronic illness and promote holistic, healthy aging.

1.1. Tailored physical activity health coaching as an effective intervention to increase PA.

One such intervention is physical activity-focused health coaching (PAHC); an abstract term defining the practice of health coaches building relationships with clients to give knowledge, foster accountability, evoke motivation for behavior change, set goals, and realize achievements relating to the goal of increasing physical activity and decreasing sedentary behavior. There are many approaches to PAHC, making generalization difficult, however, in the aggregate, PAHC promotes PA in older adults[3–5]. The efficacy of PAHC to promote behavior change is mediated by communication medium, theoretical approach, and duration of coaching[28]. In general, approaches that are tailored to individual needs, personal preferences, and outcomes of interest have a greater likelihood of promoting PA[29]. PAHC interventions must also consider the context of target populations, e.g., the needs and barriers of older adults will be different than other younger adults. Personalized PAHC can significantly increase PA[3,4], however current models of health coaching rely on the labor of health coaches – which is costly and has limited scalability. AI PAHC systems do exist, with some promising results[5,30,31], but are not technologically mature to replace health coaches. *Key limitations include the lack of adequate health-focused data to adapt complex models to health coaching applications*[5,8,30,32,33] *and the difficulty of reproducing sophisticated conversational ability and thus, meaningful tailoring of health coaching interventions*[8,30]*.* Despite its current limitations, AI PAHC is a promising technology that could increase access to PAHC for more people.

2. Innovation.

<u>The core innovations of this project address the key limitations of AI PAHC while balancing the novel challenges of this new approach: 1) the use of LLMs fine-tuned with a PAHC message-set to generate conversationally flexible and human-like responses tailored to the needs of older adults, and 2) application of the LLM-Augmenter framework to introduce external knowledge and minimize factual inaccuracy inherent to LLMs.</u> Multiple systematic reviews on health-focused chatbots and conversational agents[30,32,33] suggest that conversational flexibility and the ability to reply naturally to unconstrained input are crucial to a chatbot's ability to engage users and maintain engagement over longer periods of time. To accurately replicate the conversational flexibility, sophistication of responses, and personalization of human PAHC, we must use LLMs instead of previous constrained/rule-based systems with reduced flexibility[8]. LLMs can better handle a wide range of inputs and provide more sophisticated responses than rule-based or scripted chatbot systems[8]. But these models require large training datasets that are difficult and expensive to obtain in health-related applications like PAHC[8,9,34]. Additionally, the flexibility of these models can lead to the delivery of incorrect information, deemed 'hallucinations.' In this proposal we address these challenges, enabling the use of LLMs to provide high-quality PAHC responses.

2.1. Using of LLMs fine-tuned with a PAHC message-set to generate conversationally flexible and human-like responses tailored to the needs of older adults.

Health coaches cite basing responses on clients' lived experiences, rather than just behavioral information, as essential to successful health coaching[35,36]. As detailed above, LLMs can produce sophisticated responses to a variety of inputs, allowing them to dynamically respond to messages using information from previous messages and the shared context of the conversation. However, fine-tuning is necessary to adapt LLMs to the specific task of PAHC. Additionally, to be effective for older adults, fine-tuning must integrate the unique features relevant to aging, such as age-appropriate activity suggestions and goal setting[37]. To adapt LLMs to the specific task of PAHC for older adults, I will fine-tune a LLM with messages between health coaches and older adults collected as part of a clinical trial investigating the efficacy of personalized, motivational interviewing-based PAHC on non-sedentary behavior of adults >60. The intervention that generated these messages was based on motivational interviewing (MI) and successfully decreased sedentary behavior by around an hour per day[38] in participants who received the PAHC intervention. The 23 participants in our health coaching message set had a mean age of 72, were 56% female, and 26% Black or Asian. Fine-tuning these messages will make the LLM respond with a similar tone, style, and information content as the health coaches. This process transfers the MI theoretical approach and the older adult-focused nature of the conversations used by the health coaches to tailor their coaching and successfully promote behavior change. The use of PAHC messages from coaches and the target population, as well as the sophistication and humanness of LLM responses address the key limitations present in previous AI PAHC systems.

2.2. Integrating external knowledge and reducing hallucination inherent to LLMs with the LLM-Augmenter framework.

LLMs are not without their downsides; the same flexibility that produces dynamic responses to any input can also produce factual inaccuracies. Deemed ‘hallucinations,’ these inaccuracies represent a serious threat to patient safety, limiting the applicability of LLMs in health-focused applications. Additionally, LLMs generate responses based on patterns in the data they were trained on. When specific information is needed - such as medical information – LLMs have no way of attributing that information to a specific source. To reduce hallucination and incorporate verified external knowledge into the responses of the LLM, this project will use the LLM-Augmenter framework[11]. This framework employs a set of interconnected models to iterate through responses, determining what information is needed to answer accurately, and searching for that information in a curated document library. In a previous implementation, a system using an LLM-Augmenter eliminated hallucination during a medical Q&A task[13]. Despite these encouraging results, to prioritize patient safety I will be implementing a hybrid, human-in-the-loop PAHC system. In this system, human coaches will be supplied candidate messages from the AI system to edit and send, potentially reducing the labor of PAHC thus enabling a single coach to handle more clients or provide a greater frequency of contact. With continued future development, this system will evolve greater automation capabilities as it accumulates more conversation data.

<u>3. Approach.</u>

This project proposes to build (**Aim 1**) and evaluate (**Aim 2**) a human-in-the-loop message-based AI PAHC chat system, Companion2 – named after the initial pilot project Companion that served as a data collection testbed to generate a health coaching message dataset and test PAHC approaches[10]. Companion2 integrates an LLM fine-tuned on messages from a PAHC intervention with older adults with an LLM-Augmenter framework to generate high-quality, tailored responses to older adults for health coaches to edit and send; thereby potentially increasing the scalability of human PAHC and reach of a single health coach. Evaluation will consist of, 1) a metric-based approach comparing the meaning similarity of AI-generated health coaching responses to those from a reserved testing portion of our PAHC message set, and 2) human health coach rating of candidate messages and qualitative usability interviews with health coaches on the use of the hybrid messaging system.

3.1. Aim 1A: Replicating human PAHC responses focused on older adults with Companion2.

To achieve this, I will pair the LLM-Augmenter framework with an LLM fine-tuned with the PAHC message set described earlier. In the abstract, the LLM-Augmenter framework detailed in Peng et al. 2023[11] constrains the context of an LLM to only information from curated sources, thus integrating external knowledge into its responses while reducing hallucination. In a previous implementation, a system using an LLM-Augmenter eliminated hallucination in the model’s responses during a medical Q&A task[13]. The augmenter framework uses a system of linked text-to-text models (augmenter) to interface with a static LLM. It works in cycles – a requester submits a query (such as a client’s response in a health coach/client conversation), the LLM-Augmenter attaches external knowledge to the query, asks the static model for a candidate response, and evaluates the candidate response. This cycle of appending evidence, gathering a candidate response, and evaluating the candidate response continues until a quality threshold is reached and the final response is submitted back to the requester. Peng et al. 2023 tested this framework on two scenarios - open-ended question answering and task-oriented dialog. In both scenarios, the augmented system maintained fluency and informativeness of its responses while significantly reducing hallucination when compared to an unaugmented static LLM[11].

3.2. Building Companion2 with the LLM-Augmenter framework.

The LLM-Augmenter framework links a static LLM with the augmenter – a set of specialized sub-models, with access to a shared working memory database[11]. The LLM-Augmenter system has five components: the <u>policy module</u>, <u>utility module</u>, <u>action executor module</u>, <u>the static LLM</u>, and the <u>working memory module</u>. Each component performs a specific function.

- Queries from the coaching client are evaluated by the **<u>policy module</u>**, that can 1) direct the action executor to add external knowledge to the client input, 2) direct the action executor to generate a candidate response by submitting a request to the static LLM, or 3) release a response that has been favorably evaluated by the utility module to the overseeing health coach. The policy module and utility module are text2text neural networks that require adaptation via transfer learning to specific applications.

- The **utility module** provides common language feedback, which is integrated into the prompt used to generate the next candidate response. Feedback is based on the consistency with the evidence pulled from the knowledge retriever and the fluency of the response given the input; similar to Langchain's LLMCheckerChain, which asks the LLM to list assumptions made to answer a question, and if those assumptions are correct[39] in the context of the question. Each candidate response from the static LLM is evaluated by the utility module before being returned to the policy module.
- The **action executor** consists of an external knowledge retriever and a prompt engine. The external knowledge retriever uses an embeddings model with a vector database, specifically the Sentence-BERT embeddings model[40] with Chroma vector database[41]. This setup furnishes documents most highly related to a query from a curated document store. This project's document store will consist of documents related to 1) statistical information and health promotional materials, e.g., 2018 HHS Physical Activity Guidelines for Americans, 2) activity-related informational materials, 3) practice-based informational resources, e.g., the Open Textbook Library's Essentials of Exercise and Sport Psychology: Open Access Textbook, and 4) theoretical behavior change informational materials, e.g., Motivational Interviewing for Nutrition and Fitness[42]. These documents contain accurate external information that a health coach would use to respond to a client, such as benefits of PA, recommendations and guidelines for PA, knowledge of how to perform specific exercises, and scientific information about physical and mental processes relating to PA and behavior change. To eliminate health-related misinformation, these documents will be sourced from topic experts such as government organizations, national medical academies, regulatory organizations, and university programs. Once relevant external knowledge is obtained, the knowledge retriever summarizes the relevant evidence. The prompt engine combines the client input with the conversation history and, if either exist, any evidence generated by the knowledge retriever and any feedback from the utility module generated from the previous candidate response.
- The **static LLM** is a general-purpose, instruction pre-trained LLM. Open-source models like Llama 2 can be further fine-tuned with an appropriate training dataset to cement basal characteristics of its responses, such as tone of responses, common language, reading level, style and focus of response. I will use Meta's Llama 2's 13b parameter model[12] as has a large context length and its existing question-answering capabilities will reduce the needed size of the fine-tuning dataset[43]. Llama 2 will be further fine-tuned with 20k PAHC messages collected from the original Companion project between health coaches and coaching participants[10]. The size of this dataset is roughly similar to those used in fine-tuning of other chat-focused LLMs[44,45]. This will confer the health coach's tone, general knowledge, coaching approach, and other messaging characteristics to Llama 2.
- The **working memory module** stores conversation history, previous inputs, every candidate's response with utility module feedback and metric scores, and diagnostic information about model performance. The working memory module is accessible from all other modules.

Since each module performs a specific function, each can be updated and tested separately. This method confers greater troubleshooting flexibility as it isolates components of the response. If conversations are lacking in external knowledge, more documents can be added to the external knowledge cache, or models with different embeddings can be tested. As newer LLMs are released, I can swap static LLMs to evaluate differences in performance and response quality.

3.3. Key Challenges and Potential Solutions Relating to Implementation of the LLM-Augmenter Framework.

This approach is not without its challenges, specifically the implementation of the text2text models that power the utility and policy modules and the protection of participant information in the PAHC message set. To these challenges, I have devised potential solutions. Peng et al. 2023 used a pre-trained T5 model from Google[46] with simulated conversations from If-Then statements generated by domain experts for the specific application (e.g., customer service chatbot) for the utility and policy modules. The use of transfer learning with the pre-trained T5 model drastically reduced the size of the training dataset needed to adapt T5 as the policy module but took trial and error to develop a comprehensive model. Another possible approach would be to use the same static LLM with a chain-of-thought prompting system – a system that uses prompts to walk the model through intermediate steps of reasoning - such as LangChain. The use of chain-of-thought has been shown to elicit reasoning and self-correction in LLMs[47]. I plan to implement and test both approaches (Peng et al. 2003 and Langchain) to determine which to continue using. Implementation of the policy and utility modules will be the major challenge of Aim 1A. Another concern is the ethical implication of using health coach-client conversations from Companion to train Companion2. The initial Companion study obtained informed consent from study participants to collect their conversations to improve future interventions. Conversations have already been deidentified – identifiers (names, addresses, phone numbers, etc.) changed or removed. This, combined with the fact that Llama 2 is trained on billions of documents from across the internet, will obfuscate any information that could be used to reverse-identify the original subjects. Additionally, all training, storage, and operation of the models will be done on University hardware, limiting the exposure of training data or model weights to external systems. Further information on the protection of human subjects, and the data management and sharing plan can be found in the PHS supplemental section.

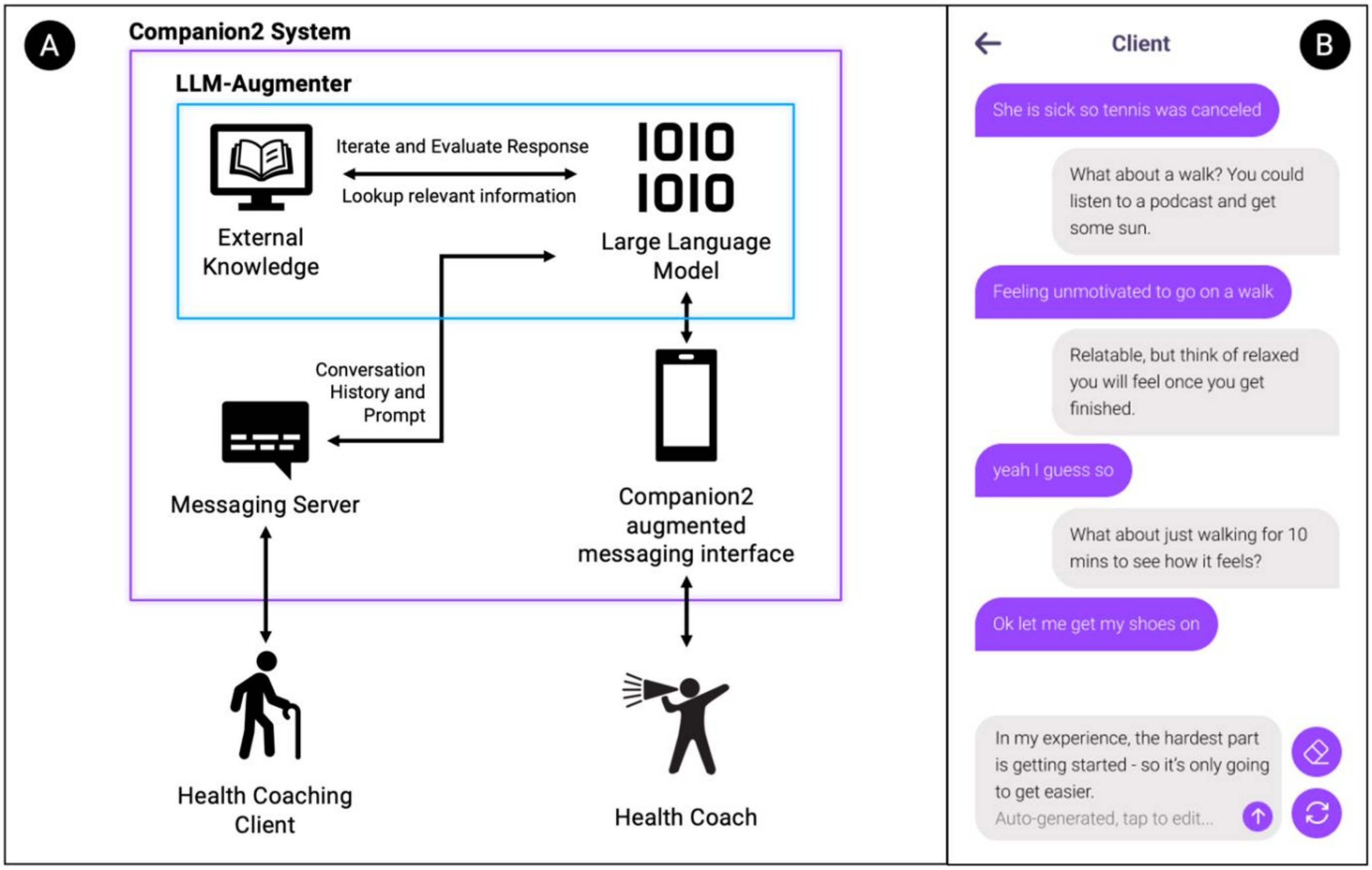

Figure 1A: Overview of the Companion2 system, **1B:** Health coach messaging user interface mockup.

3.4. Aim 1B: Implementation of a Backend and UI to Access Companion2.

Health coaches message participants over a variety of media such as WhatsApp, Signal, or SMS. Coaches can review conversation histories, message multiple participants, and access their messages from different applications and devices. These systems require a front-end user interface and a back-end system to process requests made to Companion2 and store data. In addition, a fully functional system requires a messaging provider to receive and send messages and a database to store information. (**See Figure 1A.**) Technological development of these components over the past 20 years has led to an abundance of easy to implement templates and low-or-no-cost messaging services. As this project will only require a few users, our laboratory computer can be the backend and UI server. I will use web-based frameworks like React w to create a messaging UI that is web accessible from any device, while the backend will run on Python's Django framework. Database functionality will be provided by MySQL and messaging with the participant will use Telegram's Bot API. For the scale of this project, all these components and services are free and will be configured to use end-to-end encryption.

3.5. Description of UI to Access Companion2.

The primary focus of Aim 1B is the development of a messaging UI for health coaches that supplies them with candidate responses. All health coaches have prior experience with messaging applications. Most messaging applications look relatively similar; messages stacked in ascending order by time of receipt, with the users' messages on the right and a response box at the bottom of the screen. The goal of the system UI is to integrate the candidate responses in the least intrusive manner possible – minimizing disruption of health coaches' pre-existing affordances relating to messaging apps and therefore the learning curve of the user interface[48]. To achieve this and encourage the health coach to engage with the candidate message, the candidate response will be auto-filled into the response box. If a coach is actively typing or editing a response as a new message arrives, the system does not override their work. Two buttons relating to the candidate response will always be available to the health coach – "clear" (to clear the response box) and "regenerate" (resubmit conversation history to Companion2 and regenerate candidate response.) (**See Figure 1B**)

3.6. Aim 2: Evaluation of Companion2.

The evaluation scope of this is limited to the quality of Companion2's responses; ***not*** *Companion2's ability to maintain user engagement, efficacy at producing behavior change, or ability to increase health coach productivity.* Per Abd-Alrazaq et al. 2020[49], a systematic review of technical metrics used to evaluate health chatbots, metrics-based evaluation of a system's model performance that empirically assesses response quality compared to ground truth can establish health chatbot functionality. Only after functionality is established can self-reported measures of usability and social engagement meaningfully describe a health chatbot's performance with users, and only after a system is found to be functional, engaging, and useable by coaching clients can researchers begin to ask the question of its behavior change efficacy.

3.7. Outline of the Evaluation.

Evaluation will consist of similarity scoring of Companion2 candidate responses to human responses from a reserved test PAHC message subset, computed via BERTScore[14] (Aim 2A); human evaluation of Companion2 candidate responses including subjective assessment of humanness, adherence to MI principals, and presence of hallucination; and semi-structured user experience interviews with health coaches on messaging UI (Aim 2B). Evaluation of Companion2 starts with generating candidate responses from actual message histories. To accomplish this, around 2% (~350 messages or ~50 conversation histories 5-10 messages long) of the health coach-client PAHC conversation dataset will be randomly selected and reserved as evaluation data. Companion2 will generate candidate response messages from conversation histories submitted as input. These candidate messages will then be evaluated with BERTScore and rated by health coaches. Insights gained from the initial evaluation will be used to improve Companion2 for future intervention efficacy and usability testing; to determine if Companion2 is an effective tool to increase the reach of health coaching.

3.8. Aim 2A: Evaluation of Candidate Responses Using Reserved Test Conversations with Metrics.

Companion2's candidate responses will be compared to human health coach responses using BERTScore. There are many ways to measure the performance of a health chatbot; one survey found 26 different metrics used to measure performance of research health chatbot systems[50]. With so many possible metrics, the resulting lack of standardization complicates comparison between systems[49], therefore my goal is to use metrics that are shared among other similar implementations. BERTscore[14] compares the meaning similarity between two texts and has been extensively used in evaluation of medical chatbots[51,52]. It is often used to evaluate translation algorithms, as meaning similarity is more important for language translation than word similarity. BERTscore will compare the meaning similarity between the health coach's original message from the conversation evaluation message set and Companion2's response to the same conversation history. No two messages will ever be the same, even if the AI-generated message is from an LLM trained on the health coach's conversation; and nor should they be, as there are many valid ways to convey the same information and response diversity promotes engagement. However, the messages should be similar in meaning. The similarity (BERTscore) between Companion2 and health coach responses' is arbitrary unless it can be compared to control conditions. Therefore, I will also test the meaning similarity (BERTscore) of the static fine-tuned Llama 2 model (equivalent to just Companion2's static LLM fine-tuned on health coach conversations, without the LLM-Augmenter), and the off-the-shelf Llama 2 model. By comparing the BERTscores of each of these models (Companion2, conversation fine-tuned LLM, base LLM) when attempting to replicate a human health coach's response to a conversation history, I will isolate the performance improvements conferred by each component of Companion2. I hypothesize that the average BERTscore of Companion2's responses will be higher than that of either the fine-tuned LLM or base LLM. To assess statistical significance, I will conduct an ANOVA test between the BERTScores generated from the reserved test 50 conversation histories of each experimental condition (Companion2 vs human BERTScores, conversation fine-tuned LLM vs human BERTScores, base LLM vs human BERTScores). Significance will be set to $\alpha=0.05$ and with base LLM vs human BERTScores as the reference group.

3.9. Aim 2B: Human Evaluation and Qualitative Usability Study.

Aside from automated metrics, human evaluation of Companion2 will consist of health coach (N=5) ratings of candidate message humanness, adherence to MI principles, and presence of hallucination, as well as a health coach user experience interview. Five health coaches with backgrounds in MI be recruited from Northeastern University's exercise physiology community – to which I have existing professional connections. After consent, coaches will be asked to review 50 messages and their conversation histories (up to 9 previous messages), rating each message's humanness (7 item Likert scale, not at all human-like to indistinguishable from human), adherence to MI principles (using 7 item Likert scales from the Motivational Interviewing Treatment Integrity 2.0 coding system[53,54], specifically the two global summary items, *empathy* and *MI spirit* as used in previous research[55]), and the degree of hallucination present in the message (7 item Likert scale, no hallucination present to clearly hallucination). Half of the messages to evaluate will be the original response from the health coach in the intervention (control) and half will be the Companion2 candidate response (experimental). Coaches will be blind to the origin of the message and the order of messages will be randomized for each coach. I chose the specific number of coaches to recruit - five - to mitigate the bias of any one coach when evaluating subjective measures while balancing the difficulty of recruiting practicing health coaches with MI expertise. To determine the difference between the humanness, MI adherence, and hallucination of Companion2 versus the original health coach message, we will model each outcome with a mixed-effects model with the coach as the random effect and the message origin (Companion2 or original health coach response) as the fixed effect. Depending on the variability between coaches and the mean differences between experimental conditions (both unknown) power simulations with the R package *simr*[56] indicate 50 messages should be sufficient for 80% power at α=0.05. Finally, health coaches will be given access to a prototype of the Companion2 application, complete with simulated conversations, and asked to perform specific tasks – read a conversation history, switch between conversations, edit a candidate message, and send a response. The coaches will then be interviewed in a semi-structured manner on the usability of the messaging application (ease of use, barriers to usage, etc.), quality of responses (characteristics of response, quality of the health coach edited responses, fluency of the conversation), and general wants and recommendations based on their use of Companion2. Interviews will be transcribed and analyzed for themes. This research will be conducted with approval from the University's IRB and the health coaches will be compensated for their time with a $50 gift card.

3.10. Future Directions.

The scope of this project is limited to the creation and initial evaluation of Companion2; future work will focus on usability and efficacy testing with health coaches and older adults to determine if the system can increase daily PA and health coach productivity. Future technical changes to Companion2 will include expansion of automation, such as automation of routine responses, like social conversation and basic question asking. Using policy rules implemented in the policy module, the system will determine what messages need to be elevated for a human health coach to review. Additionally, the future system will include integration of real-time PA data from wearables to better inform PAHC and provide just-in-time responses.

Respective Contributions

Prior Work

This project is based on prior work and data from the Companion project conducted by Dr. John (primary sponsor and advisor), Dr. Arguello (referee), and me. Dr. Intille (sponsor) participated in the design of Companion and provided technical support. The Companion project explored the impact of real-time activity-aware health coaching on the sedentary behavior of older adults. The text conversations generated from Companion form the basis of the fine-tuning dataset for this project.

Conceiving the Project

This project is the logical next step to Companion and was born from discussions with two of my sponsors – Drs. John and Intille. During Companion, we recognized the utility that partial automation of responses could bring to the productivity of individual health coaches. We also recognized the advances in language models combined with our amassed dataset of health coach-client messages could address the barriers of previous AI health coaching systems. Although I did not have much experience with language models, I was interested in pursuing this work for my dissertation. Dr. Intille introduced me to Dr. Wang, who was integral to the technical development of the AI system design. With their guidance and feedback, I designed the aims for this project and the research strategy.

Executing the Project

I will conduct and manage all aspects of the project. During the project, I will be co-mentored by Drs. John, Intille, and Wang. Dr. John will primarily oversee my research. For Specific Aims 1A and 1B, Dr. Wang will provide feedback and guidance on the development of the AI model and the backend system. Drs. John and Intille will provide feedback and guidance on the evaluation of the model (Specific Aims 2A and 2B). Dr. John will assist with the recruitment of health coaches from previous work. Dr. Leeman will provide guidence and advice on theoretical integration of behavior change frameworks during system construction (Specific Aims 1A and 1B) as well as guidence on evaluation of theoretical adherence during testing (Specific Aims 2A and 2B). All sponsors will assist and provide feedback on the dissemination of my work. I plan to meet with my primary sponsor and advisor, Dr. John, weekly to retrospectively examine my progress and modify tasks for the following week, as necessary. I will meet with Dr. Intille, Dr. Wang, and Dr. Leeman monthly during the completion of the specific aims. The high frequency of contact with my sponsor and co-sponsors will ensure constant feedback and monitoring of my progress towards achieving the objectives of my research project and enable necessary modifications to improve my project.

<u>Selection of Sponsor and Institution</u>

I applied to Northeastern University's doctoral program in Personal Health Informatics because of its interdisciplinary focus on health and computing. The equal emphasis placed on understanding the causes and systems that affect one's health and the experience using computing tools that can measure and intervene in health-related problems was appealing as I have a formal background in biology and many years of experience programming. Furthermore, the experience of being a graduate student at Northeastern University is best described by a fellow student I interviewed with during the admissions process. He described his doctorate at Northeastern as "a professional experience" rather than schooling, and that he was a respected contributor whose opinion was taken seriously by his advisor and collogues. Three years into my PhD in Personal Health Informatics, studying under the same advisor and my sponsor for this project, ***<u>Dr. Dinesh John</u>*****,** my experience has been consistent. Dr. John is an Associate Professor in the Department of Health Sciences at Northeastern University. When I joined Dr. John's lab, I immediately began work on an ongoing NIH-funded clinical trial studying the efficacy of activity-aware health coaching on sedentary behavior of older adults. I worked closely with Dr. John on this project and have enjoyed Dr. John's approach in training graduate students. He has given me structured autonomy to experiment with new solutions to issues we had with computation of activity classifications from accelerometry data, follow my scientific intuition, and determine what classes I needed to take to strategically advance my knowledge base to enhance my research capabilities. Dr. John's guidance on when to focus on the bigger picture, how to become a better writer, and how I should disseminate my work to build my career has resulted in me co-authoring a manuscript in *Sensors*, and presenting a poster at the *International Conference on Ambulatory Monitoring of Physical Activity and Movement*.

Northeastern University's research is centered on practical, experience-driven exploration; fostering scientific breakthroughs that directly address pertinent societal challenges, particularly in the fields of health and computing. I believe that Northeastern University's environment of collaboration led to the circumstances by which I came to know my two co-sponsors, Dr. Stephen Intille and Dr. Dakuo Wang. I met ***<u>Dr. Stephen Intille</u>***, a Professor with a joint appointment in the Khoury College of Computer Sciences and Bouvé College of Health Sciences, and the director of the Northeastern University mHealth lab, while I was working on the Companion project - an NIH funded clinical trial on the effect of activity aware personalized health coaching on sedentary older adults. As much of Dr. Intille's work involves using machine learning to understand and quantify health-related behavior to effect positive health outcomes, he was directly involved in the planning and ideation of Companion; providing the algorithms we used to classify behavior from wrist-worn accelerometers. A quality about Dr. Intille that stood out to me was his emphasis on 'practical' and intuitive solutions, and his focus on the process of iterative design and evaluation that is needed to identify and create those solutions. Dr. Intille has a history of such initiatives funded by the National Institutes of Health and the National Science Foundation. Most recently his work on the Temporal Effects on Movement and Exercise (TIME) study used micro-ecological momentary assessment and smartphone-based sensor monitoring to better understand individual variation of activity with the goal of building more predictive theories of health behavior. He has been an involved member of my doctoral studies since I matriculated and will continue to support my work as I progress through the program.

<u>Dr. Dakuo Wang</u> is an Associate Professor jointly appointed in the Khoury College of Computer Sciences and the College of Arts, Media, and Design. I had the privilege of working with him and one of his students over the summer on a project examining how smart-speaker based voice assistants can facilitate information sharing between patients and their clinical care team. Dr. Wang's welcoming and collaborative spirit, and our shared interest in leveraging large language models to solve health-related problems have sparked engaging discussions and helped me understand the complexities of this project. Our collaboration so far has showcased his commitment to teamwork and the mutual benefit of our professional relationship. I look forward to working with him more closely.

The guidance, training, and resources available from my sponsors - Dr. Dinesh John, Dr. Stephen Intille, and Dr. Dakuo Wang - as well as the collaborative environment of Northeastern University's research community, have been critical to my success in my program thus far, and will drive the success of this project, my scientific training, and future career.

Responsible Conduct of Research

My formal training in Responsible Conduct of Research (RCR) will be supplemented with informal instruction through mentoring by my co-sponsors during my doctoral training at Northeastern University. Northeastern University is committed to professional practices of scholarly and research integrity that promote and provide instruction in RCR for all faculty, students, and staff.

I completed the *Human Subjects Research for Social Behavioral Scientists Refresher Course* offered by the Collaborative Instructional Training Initiative (CITI) in August 2023. Completion of these courses is required by Northeastern University prior to engaging in human subjects research. The following modules were completed as part of the *Human Subjects Research for Social Behavioral Scientists Refresher Course* include:

Refresher Course 1 – History and Ethical Principles
Refresher Course 1 – Informed Consent
Refresher Course 1 – Defining Research with Human Subjects
Refresher Course 1 – The Federal Regulations
Refresher Course 1 – Assessing Risk
Refresher Course 1 – Informed Consent
Refresher Course 1 – Privacy and Confidentiality
Refresher Course 1 – Unanticipated Problems and Reporting Requirements in Social and Behavioral Research

I will continue to engage in refresher courses on RCR, as necessary to maintain my CITI certification. In addition, my training in RCR will continue with informal instruction throughout my doctoral education. I will engage in informal instruction with my advisor (Dr. Dinesh John) every week and with co-sponsors (Dr. Stephen Intille and Dr. Dakuo Wang) monthly. In addition to discussion of project progress, we will discuss various issues pertaining to data collection, management, and storage. My sponsor and co-sponsors have extensive training in the ethical conduct of research. I also plan to attend regularly scheduled research ethics seminars offered by the Northeastern University Office of Research and Finance. All together, these activities will exceed the federal requirements of at 8 hours per year of RCR training. Northeastern University Office of Research Enterprise Services operates the Responsible Conduct of Research Program. This program identifies trainees and fellows covered by NIH's RCR requirements, tracks their RCR training activities, and offers them priority registration for RCR instruction courses.

As part of the PhD program, I completed CS6350, Graduate Empirical Research Methods Course in the sprint of 2022. This in person, four-credit course covered many topics relating to responsible conduct of research, specifically: how to properly analyze experimental data; how to avoid scientific misconduct; ethics of human experimentation; informed consent; the reasons and processes whereby research institutions comply with governmental regulation, including human subject; and the history and development of scientific ethics.

Section II -- Sponsor and Co-Sponsor Information.

1. Research Support Available

Dr. Dinesh John

Funding Source/ID Number	PI/PD Name, Project Title	Start Date – End Date pro ected	Amount

[redacted]

Dr. Stephen Intille

Funding Source/ID Number	PI/PD Name, Project Title	Start Date – End Date ro ected	Amount

[redacted]

Dr. Dakuo Wang

Funding Source/ID Number	PI/PD Name, Project Title	Start Date – End Date ro ected	Amount

[redacted]

2. Sponsor's/Co-Sponsor's Previous Fellows/Trainees

Dr. Dinesh John: Over the past several years at Northeastern University, Dr. John has been the primary advisor and/or mentor to several masters and doctoral students. He is currently mentoring two Doctoral students and his previous students have assumed professional positions in various clinical and research settings and a sample of students and their current affiliations are listed below:

Training period	Name	Description
2021-current	Hoan Tran	Doctoral student in Personal Health Informatics at Northeastern University. Mr. Tran is currently working on Dr. John's NIH funded project that is gathering a novel data set on 24-hour behavior to develop novel algorithms to predict various aspects of waking and sleep behavior.
2021-current	Ethan Rogers (applicant)	Doctoral student in Personal Health Informatics at Northeastern University. Mr. Rogers is currently wrapping up data analyses on a recently conducted NIH funded clinical trial on which, he developed various technology solutions that were deployed over the 2-year project. While continuing to work on data analyses from this project, he is currently supported via a teaching assistantship, which is part of the training plan for PhD students in the personal health informatics program. Commencing Fall 2024 till the completion of his PhD, he will be supported by a graduate research assistantship that is available to Dr. John, which is made possible by the University to support promising students conducting research on Health AI.
2015-2022	Diego Arguello	Dr. Arguello received a PhD in Population Health Science from Northeastern University and was mentored by Dr John during his PhD. Dr Arguello is currently an Associate Director and Clinical Development Scientist at Alexion Pharmaceuticals in Boston, MA.
2019-2021	Elise Ackermans	Ms. Ackermans received an MS in Exercise Science from Northeastern University and was mentored by Dr John. She is currently employed as a full-time Exercise Physiologist at Boston Children's Hospital in Outpatient Cardiology. In her role, she conducts cardiac stress tests for adult and pediatric patients with congenital heart disease as well as working with patients in the BCH Cardiac Fitness program.
2016-2018	Alvin Morton	Dr. Morton completed an MS degree and thesis under the mentorship of Dr. John and later completed a PhD from The University of Tennessee, Knoxville. He currently an Assistant Professor at Merrimac College.

Dr. Dakuo Wang: Dr. Wang currently advises two doctoral students and has mentored a third student who completed their doctoral program in 2022. A sample of his advises students and their current affiliations is below.

2023-current	Yuxuan Lu	Doctoral student in Computer Science at Northeastern University. Mr. Lu is currently working on Dr. Dakuo Wang's NIH funded project that is developing novel LLM-based algorithms and systems to support patient-provider communication to achieve better post-surgery recovery outcome.
2020-current	Bingsheng Arthur Yao	Doctoral student in Computer Science at Rensselaer Polytechnic Institute. Mr. Yao is mentored by Dr. Dakuo Wang since the beginning of his PhD. And he is currently working on Dr. Dakuo Wang's NSF-funded project that is developing novel LLM-based algorithms and systems for K-12 education scenarios.
2020-2022	April Yi Wang	Dr. Wang received a PhD in Computer Science from University of Michigan and was mentored by Dr. Dakuo Wang during her PhD. Dr Wang is currently an Assistant Professor at ETH Zurich in Switzerland.

Dr. Stephen Intille: Dr. Intille is currently advising seven doctoral students. At Northeastern he has been primary advisor for three doctoral students who completed their degrees. At MIT, he was primary advisor to one student who received a PhD and two postdoctoral researchers and several students in a research-based MS program. A sample of Northeastern students and their current affiliations are listed below.

2021-	Tinashe Tapera	Doctoral student in Personal Health Informatics at Northeastern University. Co-advised with Prof. Varun Mishra.
2019-	Rithika Lakshminarayanan	Doctoral student in Personal Health Informatics at Northeastern University.
2016-2024	Binod Thapa Chhetry	Received PhD in Personal Health Informatics from Northeastern University. Currently on the job market.
2014-2021	Aditya Ponnada	Received PhD in Personal Health Informatics from Northeastern University. Currently Senior Researcher at MongoDB.
2014-2021	Qu Tang	Received PhD in Personal Health Informatics from Northeastern University. Currently Senior Research Scientist at Zepp Health, Shanghai, China.

3. Training Plan, Environment, Research Facilities

Mr. Rogers' training plan was developed collaboratively by his sponsor, co-sponsors, and the applicant. Mr. Rogers' short term goals from the project include (i) building an AI system using large language models (LLMs), LLM-augmenter framework, and previous conversations between health coaches and clients that produces high-quality candidate responses for health coaches, and (ii) building a user interface and backend system that integrates candidate responses into a messaging system used by health coaches to message clients, and (iii) evaluating the quality of the candidate responses and usability of the health coaching user interface. These were carefully developed and integrated into the application and training plan so that the short-term goals serve as the foundation to achieve Mr. Rogers' long-term goal of becoming a leader in multidisciplinary team-science aimed at developing and delivering effective, stratified, and personalized intervention strategies to prevent/improve chronic health conditions. A major gap in the delivery of health interventions that harness the one-on-one personal support paradigm is scalability, which allows an interventionist to simultaneously provide individual attention/care/support to multiple individuals in need. In this application, Mr. Rogers aims to leverage and integrate cutting edge innovations in AI (i.e., LLMs) with proven behavior change theory to develop an intervention approach that can be deployed using digital technologies to several individuals, while maintaining the perception of personalized attention. This application will operationalize a necessary and key initial step when leveraging an LLM-based intervention strategy to address the unsolved problem of scaling personalized real- to near-real time digital health interventions. Each sponsor was purposefully chosen to provide Mr. Rogers with comprehensive knowledge and rigorous training in the domains of preventive health, behavior health informatics, ubiquitous computing, and human-centered AI systems. Mr. Rogers will continue to leverage the curriculum of the Personal Health Informatics Program at Northeastern University to gain multidisciplinary foundational knowledge required to develop digital technologies for preventive health, and to gain exposure to expertise in theoretical and technical knowledge necessary to develop digital health solutions through seminar experiences. Potential seminars that Mr. Rogers will attend include the Personal Health Informatics seminar (monthly) and the seminar series hosted by Northeastern University's Institute for Experiential AI (fortnightly). These seminars involve invited interactive presentations from renowned experts in digital health, AI, and behavioral science, and are open to the Northeastern University community. The PHI program and other opportunities at Northeastern University enables Mr. Rogers to be part of a diverse and collaborative group of faculty and students conducting various kinds of health informatics and digital health research, which allows him to incorporate fresh perspectives in his research.

<u>Achieving training milestones and project goals</u>: As part of his training plan, Mr. Rogers will meet with his advisor (Dr. John) weekly and co-sponsors on a bi-weekly basis one-on-one, and in a separate team meeting that will be held once a month. The goal of these meetings is to ensure that the necessary progress in achieving training objectives and project goals is achieved in a timely manner. These meetings will allow Mr. Rogers to discuss both broad learning goals and specific details of the project, discuss issues and troubleshoot them collaboratively, and discuss optimal avenues to disseminate research output. Currently, Mr. Rogers meets with Dr. John on a 1-on-1 basis each week. Currently, he interacts with Drs. Intille and Wang as required; for example, he has presented in Dr Intille's summer lab retreat and has assisted Dr. Wang in recruiting and study design for a project investigating the potential for LLMs to facilitate asynchronous provider-patient communication.

Research dissemination, networking, other training: Mr. Rogers is committed to a comprehensive approach to his professional development that includes rigorous training in addition to academic coursework. The research proposed here will help him develop advanced skills in integrating artificial intelligence, digital health, and prevention research to yield solutions that are meaningful in preventing chronic disease. The project will enable him to develop research communication and dissemination skills, and it will create opportunities to develop a network of collaborations that will be instrumental in furthering his research agenda and career. Mr. Rogers is already experienced in presenting at international scientific conferences (e.g., ICAMPAM- see biosketch), and has co-authored and authored papers that are published or soon to be under review (see biosketch). This project will allow him to actively seek new opportunities and avenues to present his work and establish himself as a promising researcher in the field of health informatics. He has already identified the Association of Computing Machinery's *User Modeling, Adaptation, and Personalization (UMAP)* and *Human Factors in Computing Systems (CHI)* as conferences to present his ongoing work, as well as *Journal of Measurement of Physical Behavior* and *Medicine & Science in Sports & Exercise* as possible journals for his research.

In addition, Mr. Rogers will assist with the data analyses of my (Dr. John) ongoing and completed studies to gain experience in analyzing high-dimensional mHealth data, including data management strategies, coding, and conceptualizing statistical models for analysis. To gain comprehensive experience on top of his existing coursework in facilitating health behavior change, Mr. Rogers will complete the Certification in Intensive Motivational Interviewing at University of Massachusetts at Dartmouth. Such course work and certification will provide Mr. Rogers with a strong understanding of the principals and practices of motivational interviewing and self-determination theory, and the knowledge necessary to integrate these behavior change techniques into his work with AI.

The training program proposed here will build on Mr. Rogers' ongoing training in personal health informatics and public health and his previous experience assisting with multiple acceptability/feasibility and efficacy trials targeting mechanisms of health behavior change (e.g., smoking cessation). With the completion of his Ph.D., Mr. Rogers intends to obtain a post-doctoral position during which he can further develop his research that will allow him to customize the research output generated from the current application and establish its technical and translational merit through well designed efficacy trials. Mr. Rogers has a clear academic/career objective that is supported by a high research acumen. As his sponsor and advisor, I and his co-sponsors aim to harness this potential via the opportunities that the award will present to Mr. Rogers and help him to transition to a career as a successful independent researcher.

Sponsor (Dr Dinesh John): Dr. John is an associate professor in the Bouvé College of Health Sciences. He is the primary mentor of Mr. Rogers and sponsor of this application. He will oversee Mr. Roger's management of the project, guide Mr. Rogers in planning, organizing, and prioritizing various tasks and components of his training, academic responsibilities, and activities to periodically disseminate research findings that will be generated during this project. Dr. John is appropriate to fulfill this role as the primary mentor to Mr. Rogers because he is a recognized expert in the field of exercise science and has extensive experience in developing and applying digital health and technology-based interventions to measure and modify physical activity and sedentary behaviors. Dr. John has served as the Program Director for the MS in Exercise Science Program, is a core faculty of the PhD in Health Informatics program at Northeastern University, and has mentored several Masters and PhD students during his 11 years as a faculty member at Northeastern University. His lab currently consists of two PhD students, two full-time undergraduate co-op students, and two undergraduate interns. Dr. John has also detailed at the National Institute on Aging (2022-2023) and conducted a portfolio analysis that is enabling the Institute to develop new initiatives in the field of digital health for aging. The experiences have enabled Dr. John to develop a tailored approach to mentoring and training individual students while ensuring broad goals of the program and student are achieved in a timely manner.

Co-sponsor 1 (Dr. Stephen Intille): Dr. Intille is a jointly appointed professor in the Khoury College of Computer Sciences and the Bouvé College of Health Sciences. He will co-mentor Mr. Rogers, providing guidance on the design, development, and evaluation of mobile health applications that support health behavior change and maintenance, especially using advanced AI and sensing. Dr. Intille has over 15 years of experience in the measurement of behavior and the use of real-time behavior measurement from mobile devices to explore new just-in-time adaptive interventions. Dr. Intille is the Program Director for the Personal Health Informatics Doctoral Program and the Area Chair for Human-Centered Computing in Khoury College. He has ongoing NIH-funded projects that focus on the longitudinal measurement of physical activity, sedentary behavior, and sleep, and on real-time interventions that use mobile sensor data to support behavior change. He currently advises five doctoral students and co-advises two additional doctoral students, each of whom works on projects that involve use of mobile sensing technology to understand and respond to behavior. Dr. Intille also collaborates closely with Dr. John on a major project designed to collect data to permit the evaluation of new algorithms for accurate detection of physical activities, sedentary behaviors, and sleep, and he has multiple students current working on the use of AI, such as deep neural networks and large-language models, for modeling behavior. Mr. Rogers will be encouraged to interact with and learn from Dr. Intille's students and to participate in group and project meetings to advance his learning.

Co-sponsor 2 (Dr. Dakuo Wang): Dr. Wang is jointly appointed associate professor in the Khoury College of Computer Sciences and the College of Arts, Media, and Design. He will co-mentor Mr. Rogers, advising and providing guidance on the design, fine-tuning, and evaluation of the language models being used in this project. As a former principal investigator at the MIT-IBM Watson AI Lab, Dr. Wang has extensive experience with human centered computing and AI. His research focuses on the challenge of transitioning health-focused AI technologies from labs to real-world clinical settings, patient perspectives on AI systems in clinical settings, and how AI systems can enhance and engage patients outside of the clinical settings. Dr. Wang currently advises 2 students. Mr. Rogers has experience working collaboratively with Dr. Wang's students and is encouraged to share and present his work with them to practice his scientific communication skills and gain feedback from his peers.

4. Number of Fellows/Trainees to be Supervised During the Fellowship:
Dr. Dinesh John: Dr. John is currently the primary mentor for 2 PhD in health Informatics students at Northeastern University, Mr. Ethan Rogers (applicant), and Mr. Hoan Duc Tran (anticipated graduation, 2026). Along with Mr. Rogers, Dr. John will continue to mentor Mr. Duc Tran during the Fellowship. Additionally, Dr John plans to enroll an additional PhD. student under his mentorship in the fall of 2025.
Dr. Stephen Intille: Dr. Intille is currently the primary mentor for five doctoral students at Northeastern University and co-mentor for two additional doctoral students. One will graduate in 2024, and he is likely to accept one new doctoral student to enter in September of 2024.
Dr. Dakuo Wang: Dr. Wang is currently the primary mentor for two doctoral students at Northeastern University and co-mentor for two additional doctoral students. One will graduate in 2024, and he is likely to accept two new doctoral student and one postdoc to enter in September of 2024.

5. Applicant's Qualifications and Potential for a Research Career.
Mr. Rogers has demonstrated the potential to become an exceptional independent investigator. In his time as a student in the Personal Health Informatics Program at Northeastern University, he has shown initiative, dedication, and a high level of teamwork. He has gained experience on multiple research teams:

- A project on development of technology solutions for a clinical trial aimed at reducing sedentary behavior in older adults where he assisted in data collection, analysis, and dissemination.
- A project developing tools that use existing heart monitoring technologies for real time monitoring of participants engaged in virtual exercise classes where he designed and constructed a prototype system.
- A project conducting evidence synthesis exploring the possibility of using wearable technologies to classify cognitive engagement, as part of a broader system to classify Alzheimer's Disease and related dementias.
- A project exploring the use of smart speakers to facilitate asynchronous communication between patients and clinicians where he assisted in development and recruitment.

Prior to his doctoral studies, Mr. Rogers developed strong methodological experience through his work on a team testing behavior change interventions targeting smoking cessation resulting in three peer-reviewed publications. In my role as his mentor over the past two years, I have found that Mr. Rogers possesses a high capacity to independently take on multiple and complex responsibilities. For example, in his first year he completed his academic course requirements earning A's well while fulfilling his responsibilities on my research team managing and analyzing data for our ongoing clinical trial, presenting his work at an international conference, and developing a manuscript for publication. The latter outlines an open-source pipeline to receive and process 24-hour sensor data for real/near-real-time physical activity interventions. Mr. Rogers demonstrates a curiosity, passion, and dedication to this work that enhances our lab's dynamic, and he has become a critical part of our research team. He is always bringing up new questions regarding methodology, demonstrates a high capacity to understand and apply knowledge, and is driven to investigate issues and problem-solve innovatively. Without reservation, I believe that Mr. Rogers' work has been instrumental in the success of his lab's recently completed clinical trial in older adults, which is the basis for the current application. Mr. Rogers is currently working on two research questions utilizing the data generated in this clinical trial: (i) using mixed effects models to better characterize the effect and variation of health coaching on types of non-sedentary activity, and (ii) investigating the temporal relationship between just-in-time messaging during activity aware health coaching and client participation in non-sedentary activity.

Mr. Rogers is on track to complete all course requirements and qualifying exam components of his doctoral program by early summer 2024, which will allow him to concentrate on the research portion of his training for the remainder of his program. His coursework and research responsibilities, along with the training proposed under this award, will enable Mr. Rogers to gain the research experience and expertise, and develop the professional network required to develop evidence-based AI/digital health technology solutions to alleviate the increasing burden of chronic disease attributable to reduced physical activity.

Mr. Rogers has the attributes to become a successful independent investigator; he demonstrates a strong attention to detail, an ability to collaborate across multiple research teams, dedication, and curiosity, and he has a focused and potentially high-impact research agenda. While Mr. Rogers is fully capable of becoming an independent investigator designing, studying, and implementing digital health technologies as integrative components of intervention aimed at improving physical activity and health, he has limited experience with the implementation of client-facing behavior change techniques, adapting LLMs to health applications, and disseminating his work to interdisciplinary audiences. The sponsor and co-sponsor team, the training plan, resources in an around Boston, and protected time outlined in this fellowship proposal will provide critical training and tools and resources in these areas to enable him to develop into a well-rounded and independent investigator. Representing Mr. Rogers' mentorship team and cosponsors of this project, I have complete confidence in his ability to complete the proposed research training activities and to contribute the highest quality research to the fields of preventive health and computer science.

Northeastern University
Bouvé College of Health Sciences

[redacted]

Northeastern University
Bouvé College of Health Sciences

[redacted]

Institutional Environment and Commitment to Training

Ethan Rogers is a fourth-year student in the Personal Health Informatics (PHI) doctoral program at Northeastern University. The PHI doctoral program is a joint program between the Khoury College of Computer Sciences and the Bouvé College of Health Sciences. With a focus on interdisciplinary research, the PHI program prepares students to discover transformative solutions to global health challenges through technology and computing. Ethan has completed his coursework and teaching requirements as a TA for a graduate statistics class, and has started the qualifying examinations. PHI qualifying examinations take the form of a research paper requirement and two area exams; Ethan has started his paper requirement and health exam, and has completed his technical exam already. Ethan will complete his remaining obligations for candidacy by the end of fall of 2024, before the potential start date of this award. Although Ethan's thesis committee has yet to be fully assembled, his sponsors - Dr. John, Dr. Intille, and Dr. Wang - will be members.

Personal Health Informatics Doctoral Program

The PHI Program is an interdisciplinary program that educates students to conduct and lead personal health informatics research, with the goal of training the future leaders of healthcare technology. In recognition of technology's impact on the structure and delivery of healthcare, PHI faculty are committed to ensuring the next generation of health technology leaders have the skills, background, and relationships necessary to conduct research evaluating person-facing health and wellness systems. PHI graduates are focused on creating innovative, data-driven solutions to problems of healthcare, health inequity, and health access and applying those changes to systems patients and people use directly.

PHI students receive a rigorous, transdisciplinary education preparing them to conduct and lead personal health informatics research. The program accepts students of varying scientific backgrounds and disciplines. Students receive technical training in software development of computer applications for health, evaluation methodologies for health technologies, iterative and user-centered design, theory and practice of health behavior change, and experimental design and statistics. Students are encouraged to take elective classes that best align with their research needs creating a unique, research-driven experience for each student. Aside from highly knowledgeable computer scientists, PHI students are trained to be precise and adaptable scientific communicators, capable of proficient technical writing, collaboration with medical practitioners, and fostering empathetic connections with patients and end-users.

Doctoral Training

Students are trained to be strong independent researchers. To achieve this, students must publish high-quality research in rigorous academic journals and direct their own research goals at the conclusion of their studies. Students learn primarily via their research, collaborations with their peers, and experiences at conferences and internships. Coursework is useful to address gaps in knowledge but is not the primary focus of the program. PHI students are expected to seek guidance from their advisors, self-direct their learning, and pursue their academic interests – all with curiosity and a strong tolerance for failure. At the culmination of their training students should be able to independently identify pertinent research questions, design experiments, conduct research, and compellingly disseminate their findings.

Key Academic Requirements

The program, which began in 2011, currently includes 10 full-time doctoral students. Students are selected who have demonstrated technical expertise, research experience, and substantial interest in personal health informatics. The timeline of graduation is different for every student, but students generally take five to six years to complete their degree.

Requirements for Graduation

- 36 credit hours of core courses in human-computer interaction, statistics, programming, research methods, and personal health technology and evaluation. 12 credit hours of electives pertinent to the student's research. 48 total credit hours of coursework at a B average or above.
- One semester of teaching (as a teaching assistant or instructor of record).
- Students must pass their qualifying examinations, consisting of three parts:
  - *Research Component:* Students must submit original research to a high-quality peer-reviewed journal where they are the lead intellectual contributor. Their advisor must endorse this work, and students must do a public presentation on their work for the PHI community.
  - *Health Exam:* Students complete a health-focused exam created by a faculty member with a health background. The format and health topic of the exam is variable.
  - *Technical Exam:* The technical exam is developed in collaboration with a faculty member. Like the health exam, the format and health topic of the exam is variable.

- Once the above requirements are satisfied students achieve candidacy and begin their dissertation. Their dissertation is composed of a proposal and defense before the student's Doctoral committee.

Advisor

Each student accepted into the PHI program will be assigned one or more mentors. In general, a student's mentor will become their official thesis advisor. Advisors may hold a faculty appointment in either the Khoury College of Computer Sciences or the Bouvé College of Health Sciences at Northeastern, so long as they are core or affiliated faculty within the PHI program.

Dissertation Committee

Comprised of three faculty, a student's dissertation committee must include their primary advisor, an additional Northeastern PHI faculty member, and an external expert in the research topic. Students may consider an extra internal or external member. Committee members should possess a Ph.D. or M.D. (or equivalent) and be actively engaged in research and publication.

Dissertation Proposal

Upon approval of the dissertation committee, students collaborate with committee members to finalize their dissertation proposal. Their proposal should outline the proposed research, relevant literature, research problems, techniques, pilot data, and a realistic schedule. Students present their proposal, limited to around fifteen pages in conference paper format and 40 minutes of presentation time, for an oral defense before the dissertation committee. This process allows students the opportunity to practice defending their work, and to integrate their advisors' feedback, strengthening the research.

Dissertation Defense

The capstone of the PHI doctoral program is the student's completion and defense of their dissertation. Their dissertation, including the manuscript and related publications, along with the oral presentation, should demonstrate the candidate's ability to conduct independent research. The research described should meet publication standards in a reputable journal or conference. Throughout the dissertation process, students work closely with their Dissertation Committee, holding regular meetings and sharing results. During the defense, students address questions from their dissertation committee and research community members about their work, related work, and future directions.

Estimated Timeline

- **Year 1**
  - *Core Courses*. Students begin the program by taking introductory core courses on health behavior change and personal health technology.
  - *Research.* Students should begin integrating themselves into their advisor's work, attending seminars, and exploring their academic interests.
- **Year 2**
  - *Core and Elective Courses***.** Complete core courses and begin taking elective and readings courses.
  - *Research***.** Students should be well integrated into their research and actively working on their research paper requirements.
  - *Candidacy*. Students should begin satisfying the requirements of the three qualifying exam components.
- **Year 3**
  - *Complete remaining courses/electives*. Students can take additional courses beyond year 3, if agreed to by their advisor.
  - *Candidacy*: Students should achieve candidacy by completing their comprehensive exams and begin their dissertation. Students should complete their teaching requirements.
  - *Research*, including submission of papers to peer-reviewed journals and active dissemination at conferences.
- **Year 4**
  - *Complete remaining courses/electives*. Students can take additional courses beyond year 3, if agreed to by their advisor.
  - *Continue research*, including submission of papers to peer-reviewed journals and active dissemination at conferences.
  - *Dissertation Proposal*. Students should complete and defend their dissertation proposal and begin their dissertation work.
- **Year 5**
  - *Dissertation Defense*. Students should complete the dissertation and pass the dissertation defense. At this point, students should be actively seeking their next opportunity.

Prepared by: Ethan Rogers in consultation with Dr. Stephen Intille. Edited by Stephen Intille.

PHS Human Subjects and Clinical Trials Information

OMB Number: 0925-0001

Expiration Date: 01/31/2026

Use of Human Specimens and/or Data

Does any of the proposed research in the application involve human specimens and/or data *	❍ Yes	● No
Provide an explanation for any use of human specimens and/or data not considered to be human subjects research.		
Are Human Subjects Involved	● Yes	❍ No
Is the Project Exempt from Federal regulations?	● Yes	❍ No

Exemption Number: ❑ 1 ☑ 2 ❑ 3 ☑ 4 ❑ 5 ❑ 6 ❑ 7 ❑ 8

Other Requested Information

Human Subject Studies

Study#	Study Title	Clinical Trial?
1	Building and evaluating a scalable health coaching system with large language models to promote physical activity in older adults.	No

Section 1 - Basic Information (Study 1)

1.1. Study Title *

Building and evaluating a scalable health coaching system with large language models to promote physical activity in older adults.

1.2. Is this study exempt from Federal Regulations *	● Yes	❍ No

1.3. Exemption Number ❑ 1 ☑ 2 ❑ 3 ☑ 4 ❑ 5 ❑ 6 ❑ 7 ❑ 8

1.4. Clinical Trial Questionnaire *

1.4.a. Does the study involve human participants?	● Yes	❍ No
1.4.b. Are the participants prospectively assigned to an intervention?	❍ Yes	● No
1.4.c. Is the study designed to evaluate the effect of the intervention on the participants?	❍ Yes	● No
1.4.d. Is the effect that will be evaluated a health-related biomedical or behavioral outcome?	❍ Yes	● No

1.5. Provide the ClinicalTrials.gov Identifier (e.g. NCT87654321) for this trial, if applicable

Section 2 - Study Population Characteristics (Study 1)

2.1. Conditions or Focus of Study

- Mentoring

2.2. Eligibility Criteria

Individuals with experience health coaching and a theoretical background in behavior change.

2.3. Age Limits	Min Age: 18 Years	Max Age: N/A (No limit)
2.3.a. Inclusion of Individuals Across the Lifespan	Inclusion_of_Individuals_Across_the_Lifespan.pdf	
2.4. Inclusion of Women and Minorities	Inclusion_of_Women_and_Minorities-Final.pdf	
2.5. Recruitment and Retention Plan	Recruitment_and_Retention_Plan.pdf	
2.6. Recruitment Status	Not yet recruiting	
2.7. Study Timeline		
2.8. Enrollment of First Participant	01/01/2026	Anticipated

Inclusion of Individuals Across the Lifespan

Individuals under that age of 18 (children) are not included as they will not have had time to gain health coaching experience or an educational background in theory of behavior change. There is no upper age limit.

Study personnel have experience working with adults in human subject research. The PI, Mr. Ethan Rogers, has 5+ years of research experience with adults, including older adults and people diagnosed with mental health disorders. The facility where the research will be conducted is an exercise physiology lab. This site routinely handles human subject research with multiple age groups.

INCLUSION OF WOMEN AND MINORITIES

The targeted distribution of the subjects' racial and ethnic makeup is anticipated to match the demographic distribution in the state of Massachusetts based on data obtained from United States Census conducted in 2010. According to census data, the distribution of minorities in the general population in Massachusetts is 13.1% Hispanic or Latino of any race, 9.5% black or African American, 0.5% American Indian or Alaska native, 7.7% Asian, 0.1% Native Hawaiian or other Pacific Islander, and 2.7% identified as having another race from these categories. White adults are the majority at 69.6%. We will ensure that recruitment will satisfy the NIH requirement of at least 50% women and 20% minorities.

This study will be recruiting adult health coaches with backgrounds in behavior change theory. We have existing connections to the physical activity research community at the University where the proposed research is being conducted and from where we plan to recruit. This community is diverse in gender, race, and ethnicity; we do not anticipate to needed to conduct special recruitment efforts for women and minorities.

This study will use data to train machine learning models to be tested by the recruited health coaches. These data include anonymized conversations between a health coach and 23 coaching clients, collected with consent as part of a previous clinical trial. 13 (56%) of these participants were women, and 6 (26%) were minorities.

Recruitment and Retention Plan

This usability evaluation of a health coaching messaging system with will recruit individuals with health coaching experience and educational backgrounds in theories of behavior change. Through previous research we have established connections with practicing health coaches, who we plan to recruit by word of mouth. These coaches come from a diverse physical activity research community encompassing personal trainers, personal training students, and exercise physiology researchers and educators. We will start by asking known health coaches from previous work. Once recruited, we plan to continue recruitment by asking them to disseminate our study contact information. As our sample size is very small (n=5) we believe this to be adequate. In the scenario word of mouth is inadequate, out contingency plan is to post flyers in recreational spaces (specifically gyms and recreation facilities) and request inclusion of our usability study in the weekly bulletins of health-affiliated departments.

Once recruited, the study requirements include a single hour long visit to the exercise physiology lab to rate AI health coaching responses generated from our system on a variety of scales, use a prototype interface developed by our team, and participate in a qualitative interview about the user interface. We plan to offer compensation in the form of $50 gift certificates. The minimal requirements of the study and compensation will retain participants.

2.9. Inclusion Enrollment Reports

IER ID#	Enrollment Location Type	Enrollment Location
Study 1, IER 1	Domestic	Northeastern University

OMB Number: 0925-0770

Expiration Date: 09/30/2024

Inclusion Enrollment Report 1

1. Inclusion Enrollment Report Title* : Inclusion Enrollment Report for Building and evaluating a scalable health coaching system with large language models to promote physical activity in older adults.

2. Using an Existing Dataset or Resource* : ● Yes ❍ No

3. Enrollment Location Type* : ● Domestic ❍ Foreign

4. Enrollment Country(ies): USA: UNITED STATES

5. Enrollment Location(s): Northeastern University

6. Comments: This survey and usability study will only be recruit 5 subjects, thus exact racial minorities may vary.

Planned

Racial Categories	Ethnic Categories				Total
	Not Hispanic or Latino		Hispanic or Latino		
	Female	Male	Female	Male	
American Indian/ Alaska Native	0	0	0	0	0
Asian	0	0	0	0	0
Native Hawaiian or Other Pacific Islander	0	0	0	0	0
Black or African American	1	1	0	0	2
White	1	1	1	0	3
More than One Race	0	0	0	0	0
Total	2	2	1	0	5

Cumulative (Actual)

Racial Categories	Ethnic Categories									Total
	Not Hispanic or Latino			Hispanic or Latino			Unknown/Not Reported Ethnicity			
	Female	Male	Unknown/ Not Reported	Female	Male	Unknown/ Not Reported	Female	Male	Unknown/ Not Reported	
American Indian/ Alaska Native	0	0	0	0	0	0	0	0	0	0
Asian	0	0	0	0	0	0	0	0	0	0
Native Hawaiian or Other Pacific Islander	0	0	0	0	0	0	0	0	0	0
Black or African American	0	0	0	0	0	0	0	0	0	0
White	0	0	0	0	0	0	0	0	0	0
More than One Race	0	0	0	0	0	0	0	0	0	0
Unknown or Not Reported	0	0	0	0	0	0	0	0	0	0
Total	0	0	0	0	0	0	0	0	0	0

Section 3 - Protection and Monitoring Plans (Study 1)

3.1. Protection of Human Subjects	Protection_of_Human_Subjects.pdf
3.2. Is this a multi-site study that will use the same protocol to conduct non-exempt human subjects research at more than one domestic site?	❍ Yes ● No ❍ N/A
Single IRB plan attachment	
3.3. Data and Safety Monitoring Plan	DMS.pdf
3.4. Will a Data and Safety Monitoring Board be appointed for this study?	❍ Yes ❍ No
3.5. Overall structure of the study team	

Protection of Human Subjects

Overview

This research uses human subjects in two ways, and we claim both fall under the 2018 exemptions for human subjects' research. Briefly, the research in this proposal has two aims. First, using conversations collected between health coaches and adult clients in previous work, fine-tune a large language model to generate candidate responses for health coaches via a health coach messaging interface. Second, evaluate the usability of the user interface with health coaches through qualitative interviews. Human subjects' factor into this research in two ways, 1) as subjects of interview and survey research, and 2) as secondary research information collected from previous work.

Research Activities Covered Under Exemption #2

Among other derivatives, exemption #2 covers research consisting of survey interview procedures that collect information about the subjects where their identity cannot be ascertained or the information collected would not place the subjects in criminal or civil liability or be damaging (specifically to the subjects financial standing, employability, educational advancement, or reputation). In our research we plan to conduct 1) surveys with health coaches on their subjective ratings of the humanness, adherence to theoretical principles, and degree of factual inaccuracy contained in responses generated by an artificial intelligence (AI) health coaching system, and 2) interviews with health coaches on their opinions about a prototype user interface designed to supply them with AI generated candidate responses. Although not directly of interest, this interview may touch upon their personal information (including employment, previous experience health coaching, phone habits) in a manner that, although unlikely, could potentially allow them to be indirectly identified <u>if not censored</u>. However, regardless of the censorship status of the interview or survey content, the data collected – ratings of message humanness, rating of message adherence to theoretical principles of behavior change, ratings of factual inaccuracy, thoughts, opinions, feelings about health coaching, the usability of the messaging interface, and their recommendations for future alterations – would not be damaging were subjects to be indirectly identified. As such we are claiming that this use of human subjects in our research is exempt under exemption (2.i.).

Research Activities Covered Under Exemption #4

Exemption #4 covers secondary research use of previously collected information and biospecimens for which consent is not required, provided it meets one of four sub criteria. To create a large language model capable of crafting candidate responses of adequate humanness to be useful for health coaches, it must be fine-tuned on data specific to the task at hand, in this case, conversations between health coaches and adult clients. These conversations were previously collected as part of an NIH funded clinical trial investigating the effect of activity aware personalized health coaching on sedentary behavior of older adults [Grant: 5P30AG048785-07]. As part of the data safety and monitoring plan of the previous work these conversations have been censored to remove identifiable information, including but not limited to names, address, phone numbers, and email address. The conversations do not contain any indirect identifiers (including participant ID). We believe this use meets the requirements of exemption (4.ii.), as 1) our use of the conversations is a secondary use of an existing ("off-the-shelf") dataset, 2) the conversations do not contain any directly or indirectly identifiable information that could allow subjects' identities to be readily ascertainable, 3) we will not contact the subjects, and 4) we will not re-identify the subjects. More information on the procedures we plan to use to safeguard these data, and the data generated by this research is available in the Data Management Plan.

DATA MANAGEMENT AND SHARING PLAN

Element 1: Data Type

A. Types and amount of scientific data expected to be generated in the project:

This project will generate three distinct types of data from its work.

1) *Large Language Model (LLM) Weights, stored in a serialized binary (.bin) files totaling roughly 10-50gb depending on compression scheme. At the heart of LLMs are matrices that contain the weights the model applies to each parameter. Weights are the values that produce meaningful answers from input sequences. A helpful analogy could be the arrangement of neurons in a brain; without context weights are meaningless but when loaded into the model the weights contain the information learned during training. Although the model being used in this research has publicly available weights, the alterations to the weights made during the fine-tuning process will encode the tone, style, and content of the health coaches when responding to the coaching clients.*
2) *A dataset of candidate responses (generated from the system based on input of conversation history) rated by human health coaches for each messages' humanness, relevancy, adherence to motivational interviewing principals, and degree of hallucination.* Each row will be the response from one health coach, totaling [N health coaches, projected to be 5] X [N candidate messages for rating, projected to be between 25 and 50] rows, stored in CSV format.
3) *Transcripts of semi-structured interviews with health coaches on the usability of the health coaching interface.*

B. Scientific data that will be preserved and shared, and the rationale for doing so:

The dataset of candidate responses and their ratings by health coaches will be shared to facilitate verification of the statistical analysis upon publication of this work. This will not include metadata about the coaches (to preserve their anonymity). This will include the conversation history used as input to the system to generate the candidate response. Although the conversation history messages used for fine-tuning and evaluation have been deidentified, specific details my still be captured in the message set. Thus, to preserve anonymity of the original health coaching participants, we will censor potentially identifiable information including but not limited to addresses, health conditions, physical descriptions of participants, and other unique information.

The model weights, conversation histories not used to generate candidate responses for evaluation, and full transcripts of the interviews with health coaches will be preserved for future research and evaluation but will not be shared to eliminate any threat to the anonymity.

C. Metadata, other relevant data, and associated documentation:

A data dictionary will be provided that explains the content of each variable in the released dataset, specifically the wording of any survey questions (humanness, relevancy, adherence to motivational interviewing principals, and degree of hallucination), the response options, and any notes on the collection of the data. Methods – detailed to the degree necessary to replicate this research – will be included in publication of this work.

Element 2: Related Tools, Software and/or Code:

No special tools will be needed to access the shared scientific data mentioned above.

Element 3: Standards:

Data to be shared will be stored in a single CSV file with UTF-8 encoding – an encoding shame used by all modern computers. IDs of the health coaches will be integers, assigned randomly.

Element 4: Data Preservation, Access, and Associated Timelines

A. Repository where scientific data and metadata will be archived:
As the dataset and metadata for sharing will be small (less than 2gb) the dataset will be shared as supplemental materials on PubMed Central.

B. How scientific data will be findable and identifiable:
The dataset and metadata will be shared as supplemental materials to published research, and therefore be associated with a publication and DOI.

C. When and how long the scientific data will be made available:
The data will become available at the time of publication of this research and be available for as long as PubMed Central retains submissions.

Element 5: Access, Distribution, or Reuse Considerations

A. Factors affecting subsequent access, distribution, or reuse of scientific data:
The dataset of candidate responses, the censored conversation histories used as input to the model to generate the candidate responses, and the ratings of health coaches (data shared as part of supplemental materials) will be shared without limits on access, distribution, and reuse.

As mentioned above the message dataset used to create the system, as well as the model weights that encompass the learned model, will not be shared to prevent any possible breaches of anonymity.

B. Whether access to scientific data will be controlled:
Access to the dataset of candidate responses, the censored conversation histories used as input to the model to generate the candidate responses, and the ratings of health coaches will not be controlled.

C. Protections for privacy, rights, and confidentiality of human research participants:
No identifiable information, or information that could be possibly used to reverse-identify the health coaches providing ratings of candidate messages will be provided in the dataset or the metadata. Coaches will be represented in the dataset as integers.

Although the conversation history messages used for fine-tuning and evaluation have been deidentified, specific details my still be captured in the message set. Thus, to preserve anonymity of the original health coaching participants, we will censor potentially identifiable information including but not limited to addresses, health conditions, physical descriptions of participants, and other unique information. Prior consent to use these data for this purpose was obtained.

As mentioned above the message dataset used to create the system, as well as the model weights that encompass the learned model, will not be shared to prevent any possible breaches of anonymity.

Element 6: Oversight of Data Management and Sharing:

The PI/Trainee (Ethan Rogers) will be responsible for collection, storage, and analysis of the data generated by this study. The lead sponsor of this project – Dr. Dinesh John – will ensure and verify compliance with this plan. Data will be kept on a file server with limited access; only Ethan Rogers and Dinesh John will have access to the data used to create the system and generated by the project. Dr. John will inspect the supplemental materials (mechanism of distribution) before submission to PubMed Central to ensure this plan is being followed.

Section 4 - Protocol Synopsis (Study 1)

4.1. Study Design

4.1.a. Detailed Description

4.1.b. Primary Purpose

4.1.c. Interventions

Type	Name	Description

4.1.d. Study Phase

Is this an NIH-defined Phase III Clinical Trial? ❍ Yes ❍ No

4.1.e. Intervention Model

4.1.f. Masking ❍ Yes ❍ No

❏ Participant ❏ Care Provider ❏ Investigator ❏ Outcomes Assessor

4.1.g. Allocation

4.2. Outcome Measures

Type	Name	Time Frame	Brief Description

4.3. Statistical Design and Power

4.4. Subject Participation Duration

4.5. Will the study use an FDA-regulated intervention? ❍ Yes ❍ No

4.5.a. If yes, describe the availability of Investigational Product (IP) and Investigational New Drug (IND)/ Investigational Device Exemption (IDE) status

4.6. Is this an applicable clinical trial under FDAAA? ❍ Yes ❍ No

4.7. Dissemination Plan

Delayed Onset Studies

Delayed Onset Study#	Study Title	Anticipated Clinical Trial?	Justification
The form does not have any delayed onset studies			