Enhancing Adaptive Behavioral Interventions with LLM Inference from Participant-Described States

Karine Karine KARINE@CS.UMASS.EDU
University of Massachusetts Amherst
Amherst, MA, USA

Benjamin M. Marlin MARLIN@CS.UMASS.EDU
University of Massachusetts Amherst
Amherst, MA, USA

Abstract

The use of reinforcement learning (RL) methods to support health behavior change via personalized and just-in-time adaptive interventions is of significant interest to health and behavioral science researchers focused on problems such as smoking cessation support and physical activity promotion. However, RL methods are often applied to these domains using a small collection of context variables to mitigate the significant data scarcity issues that arise from practical limitations on the design of adaptive intervention trials. In this paper, we explore an approach to significantly expanding the state space of an adaptive intervention without impacting data efficiency. The proposed approach enables intervention participants to provide natural language descriptions of aspects of their current state. It then leverages inference with pre-trained large language models (LLMs) to better align the policy of a base RL method with these state descriptions. To evaluate our method, we develop a novel physical activity intervention simulation environment that generates text-based state descriptions conditioned on latent state variables using an auxiliary LLM. We show that this approach has the potential to significantly improve the performance of online policy learning methods.

1. Introduction

The use of reinforcement learning (RL) methods (Sutton et al., 1998) to support health behavior change via personalized and just-in-time adaptive interventions is of significant interest to health and behavioral science researchers focused on problems such as smoking cessation support and physical activity promotion. (Coronato et al., 2020; Liao et al., 2020; Gönül et al., 2021; Yu et al., 2021). However, in the adaptive behavioral intervention domain, RL methods are often applied using a small collections of affective, behavioral, physiological and/or environmental context variables to mitigate the significant data scarcity issues that arise from practical limitations on adaptive intervention trials. This includes limited numbers of intervention opportunities per day, limited numbers of study participants, and limited overall study durations.

An important consequence of the use of small state spaces to mitigate data scarcity issues is that the resulting RL policies have an extremely narrow view of the overall state of an intervention participant's health and well-being. This can result in an adaptive intervention system recommending intervention options that range from sub-optimal to inappropriate

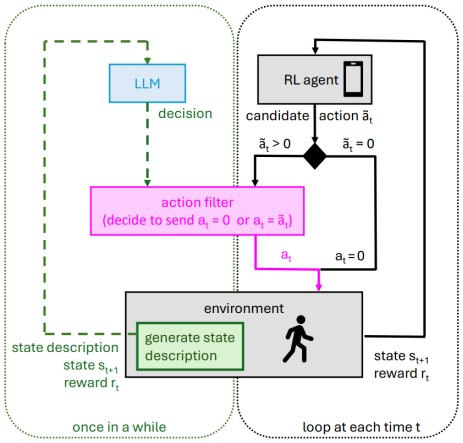

Figure 1: LLM4TS is a hybrid method that combines LLM inference with a base RL method to improve action selection. The RL agent proposes a candidate action $\tilde{a}_t$. A pre-trained LLM is then used to infer whether the action is aligned with a participant-provided state description. The LLM inference step uses a prompt with multiple components including questions that guide chain of thought-like reasoning.

with respect to the participant's overall state. For example, a physical activity adaptive intervention that conditions only on location, weather, temperature and recent activity level would have no ability to account for the fact that the participant has the flu or has sprained their ankle and cannot walk. Continuing to issue intervention content that disregards the overall health state of a participant may needlessly contribute to increasing habituation (Dimitrijević et al., 1972; Liao et al., 2018) as well as risk of disengagement from the intervention (Park and Lee, 2023).

In this paper, we propose to leverage the natural language understanding and reasoning capabilities of pre-trained large language models (LLMs) (Vaswani et al., 2017; Achiam et al., 2023; Grattafiori et al., 2024) to help mitigate the problem of restricted state spaces in RL-based adaptive interventions without impacting data efficiency or compromising the ability of behavioral science researchers to control intervention content. The approach that we explore is based on (1) enabling intervention participants to describe any aspect of their state in free text, (2) using a data-efficient RL algorithm with limited state to propose a candidate action, and (3) applying an LLM with a specifically engineered prompt to perform inference with the goal of deciding whether the proposed action is aligned with the participant's most recently declared state. We use Thompson sampling as a data-efficient base RL algorithm in this work (Russo et al., 2018; Chu et al., 2011; Thompson, 1933). We provide an overview of our method in Figure 1, which we refer as LLM4TS following the taxonomy of Pternea et al. (2024).[1]

To evaluate our approach, we build on a recently introduced simulation environment that is based on an adaptive messaging intervention for physical activity promotion. This simulation includes models of key aspects of behavioral intervention dynamics including intervention habituation and disengagement risk (Karine and Marlin, 2024). We add to this system a simulation of participants responding to the morning query about their general health state. We generate the responses using an LLM prompt that conditions on latent dimensions of the true underlying health state of the simulated participant. In this work, we focus on a latent binary state indicating whether the participant is able to engage in physical activity (specifically, walking) or not.

1. Code for LLM4TS is available at https://github.com/reml-lab/llm4ts

We present extensive results showing that LLMs can be used to reliably generate varied state descriptions when conditioned on underlying state variables, that LLMs can accurately infer when to filter actions based on state descriptions, and that the proposed LLM4TS method results in improved performance relative to the standard TS method when periods where a participant cannot engage in walking are encountered. We further explore the impact of different components of the LLM inference prompt as well as run-time metrics for the LLM inference process.

The primary contributions of this work are:

1. **LLM4TS.** We introduce an "LLM as judge" approach to enhancing personalized adaptive health interventions. LLM4TS leverages the natural language understanding and reasoning capabilities of LLMs to help mitigate the limited state representation of a Thompson sampler while maintaining data efficiency.

2. **StepCountJITAI+LLM.** We develop a novel simulation environment to evaluate the proposed method. This simulation environment extends an existing base simulator to add support for simulating participant generated text using LLMs. The simulator generates text-based descriptions of participant state that reflect latent state dimensions. This simulation environment has significant potential to enable the development of new RL algorithms tailored to the adaptive intervention domain.

Generalizable Insights about Machine Learning in the Context of Healthcare

In this work, we develop and evaluate an approach to augmenting a base RL method with LLM-based reasoning capabilities to help address the significant challenge of data scarcity that arises when applying RL methods to optimize adaptive intervention policies in the context of practical research study designs. We focus specifically on leveraging the common sense reasoning capabilities of pre-trained LLMs to better align action selection with participant-generated state descriptions. While the method is evaluated in the context of a physical activity adaptive intervention simulation in this paper, we expect that the advantages demonstrated over the base RL approach will generalize to real physical activity studies, as well as other adaptive intervention problem domains. In summary, we believe this is a promising approach for significantly augmenting the intelligence of adaptive health interventions while respecting practical constraints on study designs, retaining the ability of intervention designers to completely control intervention content, and increasing the agency of adaptive intervention study participants.

2. Background and Related Work

In this section, we describe background on adaptive interventions, intervention policy learning, simulation environments for adaptive intervention research, and other related work.

Adaptive Sequential Interventions. In an adaptive sequential intervention, the overall intervention package consists of multiple intervention *options*. Different intervention options are selected and provided to a patient or study participant at each of a collection of time points referred to as *decision points*. The selection of intervention options at each decision

point follows an intervention policy (or decision rule) that takes as input selected aspects of the overall state of the individual (Collins et al., 2007).

Adaptive sequential intervention designs have been widely studied in the behavioral science and mobile health research communities, resulting in multiple frameworks including the multiphase optimization strategy and sequential multiple assignment randomized trial (Collins et al., 2007), as well as the Just-In-Time Adaptive Intervention (JITAI) (Nahum-Shani et al., 2018). In this work, we focus on the JITAI setting, where the goal is often described as providing the right type and amount of support at the right time by adapting to an individual's changing internal and contextual state (Nahum-Shani et al., 2018).

JITAIs have been developed and studied in areas including physical activity promotion (Hardeman et al., 2019), weight loss (Forman et al., 2019), diet adherence (Goldstein et al., 2021) and tobacco and other substance use (Yang et al., 2023; Perski et al., 2022). Such interventions target health behaviors that are known to be driving risk factors for multiple chronic illnesses that account for 86% of all US healthcare spending (Holman, 2020).

Intervention Policy Learning. Given a set of intervention options, the key problem in adaptive intervention design is mapping the context (or state) of an individual into the selection of an optimal intervention option at each decision point. Since an adaptive intervention is a sequential decision making problem, the natural optimization approaches are control theory methods (Golnaraghi and Kuo, 2010) and reinforcement learning (RL) (Sutton et al., 1998). Both approaches have been used to optimize JITAIs in prior work (Liao et al., 2020; Gönül et al., 2021; El Mistiri et al., 2025). In this work, we focus on learning intervention policies using RL methods.

The correspondence between adaptive intervention terminology and standard RL terminology is straightforward: the intervention options are the actions, the participant state is the environment state, and measurements of the proximal or distal outcomes of interest forms the basis for the reward (Sutton et al., 1998). However, the adaptive intervention research study domain is a challenging setting for the application of RL methods due to practical limits on the number of decision points per day (often less than 10), the number of study participants (often 10's to low 100's), and the per-participant study duration (typically several weeks to one year). These limits result in significant data scarcity and require the application of highly data efficient RL methods.

One approach to achieve data efficiency is the application of low-variance, high-bias RL approaches such as Thompson Sampling (TS). While full RL methods attempt to estimate the future impact of present actions in each state, TS methods essentially select actions to optimize expected immediate rewards in each state. This is often referred to as the *contextual bandit* setting (Russo et al., 2018; Chu et al., 2011; Thompson, 1933). While the contextual bandit assumptions are not satisfied in the adaptive intervention optimization setting, contextual bandit methods such as TS tend to outperform full RL methods under significant data scarcity.

The standard linear Gaussian Thompson Sampling algorithm uses a reward model of the form $\mathcal{N}(r; \theta_a^\top v_t, \sigma_{Ya}^2)$, where v_t is the state vector at time t, θ_a is a vector of weights, and σ_{Ya}^2 is the reward variance for action a. Thus, $\theta_a^\top v_t$ represents the mean reward for action a. The reward model weights θ_a are treated as random variables with distribution $\mathcal{N}(\theta_a; \mu_{ta}, \Sigma_{ta})$. Actions are selected at each time t by sampling $\hat{\theta}_a$ from $\mathcal{N}(\theta_a; \mu_{ta}, \Sigma_{ta})$ for

each action a and choosing the action a with the largest value $\hat{\theta}_a^\top v_t$. The prior distribution for θ_a is of the form $\mathcal{N}(\theta_a; \mu_{0a}, \Sigma_{0a})$. The distribution over θ_a for the selected action is updated at time t based on the observed reward r_t and v_t using Bayesian inference. We provide the update equations for the mean and covariance matrix below.

$$\Sigma_{(t+1)a} = \sigma^2_{Ya} \left(v_t^\top v_t + \sigma^2_{Ya} \Sigma^{-1}_{ta} \right)^{-1} \tag{1}$$

$$\mu_{(t+1)a} = \Sigma_{(t+1)a} \left((\sigma^2_{Ya})^{-1} r_t v_t + \Sigma^{-1}_{ta} \mu_{ta} \right) \tag{2}$$

Adaptive Intervention Simulation Environments. A further challenge with the application of RL methods in the adaptive intervention setting is the extremely high cost of evaluating methods in the context of real human subjects studies. In other application areas of RL where fielding experimental methods or policies can have high cost, such as robotics, RL methods are typically developed using simulation environments (Kim et al., 2021). However, in the adaptive intervention domain, there is very limited prior work on simulation environments. In this work, we extend the physical activity adaptive intervention simulator introduced by Karine and Marlin (2024) called StepCountJITAI.

StepCountJITAI was specifically designed to support the development of new RL algorithms for the adaptive behavioral intervention domain. It simulates a messaging-based adaptive physical activity intervention. In this simulation environment, the state includes a binary context variable $c_t \in \{0, 1\}$ that can be used to model a time varying binary state such as 'stressed/not stressed' or 'at home/not at home,' etc. The simulation also models the dynamics of two key behavioral state variables: habituation level h_t and disengagement risk level d_t. The variable a_t denotes the action at time t, which corresponds to the choice of intervention option. The possible actions a_t are: do not send a message ($a_t = 0$), send an untailored message ($a_t = 1$), send a message tailored to context 0 ($a_t = 2$) and send a message tailored to context 1 ($a_t = 3$). The goal in this domain is to maximize the participant's total walking step count over the duration of the intervention. Thus, the step count at time t serves as the reward for the action taken at time t. Further details of the StepCountJITAI simulator are described in Appendix A.2. In this work, we extend the base StepCountJITAI simulator with additional state variables as well as the ability to produce natural language output describing aspects of the full state.

LLMs Combined with RL. There has been much recent work on approaches that combine Large Language Models (LLMs) with RL to address different problems (Pternea et al., 2024). The use of reinforcement learning from human feedback to fine-tune LLMs is likely the most well-known such approach (Ouyang et al., 2022), but our focus is on the opposite problem of using LLMs to enhance RL methods. Pternea et al. (2024) present a helpful taxonomy that refers to these two categories of approaches as RL4LLM and LLM4RL. The closest LLM4RL work to our approach leverages LLMs to enhance base RL policies, including work that uses an LLM as a policy prior (Pternea et al., 2024; Hu and Sadigh, 2023). Our approach falls into this same category, but instead leverages the LLM to judge proposals output by a base RL policy, an example of the LLM-as-judge framework that has previously been applied as part of the automated evaluation of models such as chat assistants (Zheng et al., 2023). To the best of our knowledge, this paper is the first to study the application of an LLM-as-judge approach in combination with a bandit-based RL method to combat significant RL data scarcity issues. Further, we evaluate LLM prompting

frameworks that leverage aspects of intermediate reasoning and provision of domain-specific knowledge to improve LLM inference (Wei et al., 2022; Lewis et al., 2020).

3. Methods

In this section, we describe our proposed method as well as our enhanced simulation environment. Figure 1 provides an overview of the proposed method.

3.1. Proposed Method: LLM4TS

LLM4TS is an adaptive intervention optimization approach that seeks to combine two sources of information about a participant's full state s_t when selecting intervention actions a_t: an observation of a limited subset of state dimensions given by $\tilde{s}_t$, and a free text description of additional aspects of state f_t provided by the participant. The problem of interest is thus how to leverage the information about s_t contained in the free text description f_t to supplement the information contained in $\tilde{s}_t$ with the goal of improving total reward relative to an approach that only makes use of the information in $\tilde{s}_t$.

For example, a physical activity intervention designer may be interested in optimizing the selection of motivational messages conditioned on a limited set of state variables $\tilde{s}_t$ including the participant's current level of stress and recent activity level. However, the participant's full state s_t includes a vast number of additional dynamic variables that may impact their ability to participate in the intervention over time. The free text description f_t provides participants with the opportunity to describe any aspect of their state s_t using natural language. For instance, using f_t, a participant could describe that they have the flu or sprained their ankle, variables not represented in $\tilde{s}_t$ that impact the ability to engage in physical activity.

Next, we note that while the collection of state descriptions f_t from participants at every decision time point t would be burdensome for interventions with multiple decision points per day, more flexible and less burdensome approaches are possible. For example, participants could be asked to respond to a query each morning asking them to describe how they are feeling or to describe any barriers to participating in the intervention. This interaction could be supplemented with the ability for a participant to provide additional descriptions via an on-demand interaction in response to events or other changes in health state. Further, while we use text directly in this work, current speech-to-text capabilities would easily allow participants to supply this information using voice, further reducing burden (Radford et al., 2023; Kuhn et al., 2024). Finally, participants may not respond to a query to provide a state description f_t at some time points t, and approaches must be able to accommodate the resulting missing data.

We propose a hybrid framework to address this problem where a base RL agent outputs a candidate action $\tilde{a}_t$ at each time step t based only on $\tilde{s}_t$. Then, based on an LLM prompt that includes the most recent participant-provided state description f_t, the LLM decides whether to allow or not allow the candidate RL action. This framework enables an end-to-end adaptive intervention where the common sense reasoning ability of a pre-trained LLM can be applied to interpret the participant-provided state description f_t to gain information about s_t that is not accessible via $\tilde{s}_t$. We describe the steps in the framework in detail below.

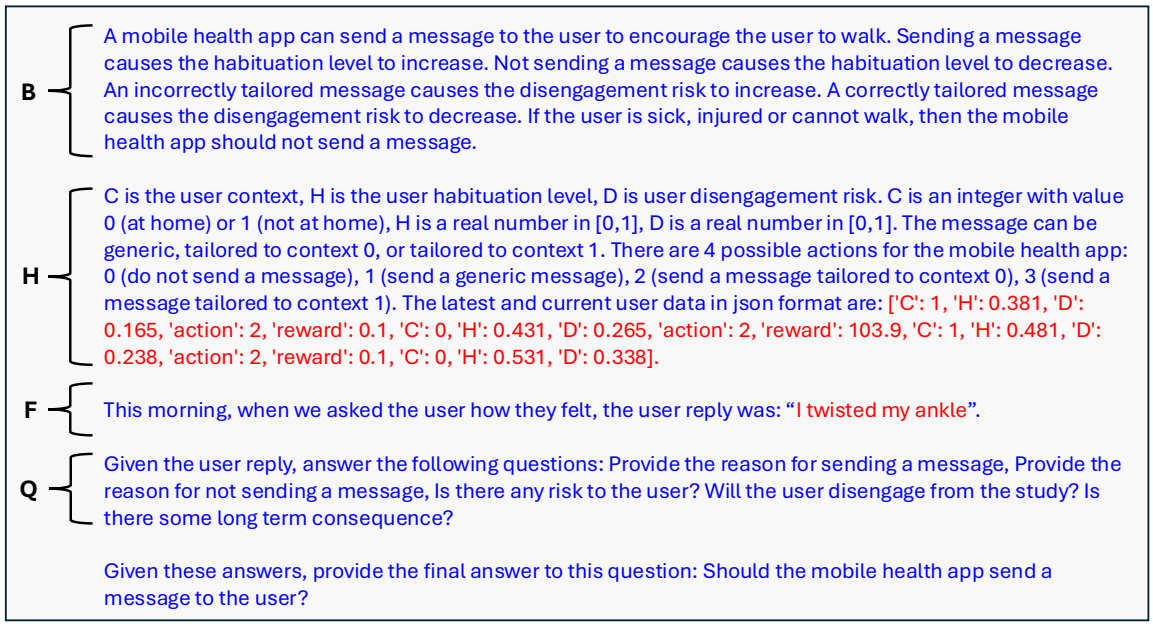

Figure 2: Example LLM prompt template with B, F, Q, H components annotated. Static prompt content is shown in blue. Dynamic prompt content is shown in red.

Without loss of generality, we assume that action 0 is a null action where no intervention content is provided to the participant.

1. **Candidate Action Generation:** At each time step t, the base RL agent proposes a candidate action $\tilde{a}_t$ based on its current parameters θ_t and the partial state observation $\tilde{s}_t$. If the candidate action is $\tilde{a}_t = 0$, the final action is $a_t = 0$. If the candidate action is $\tilde{a}_t \neq 0$, we then apply LLM inference.
2. **LLM Inference:** Given the most recent participant provided state description f_t and a prompt template, we construct a specific LLM prompt and apply an LLM to perform inference. We then extract the LLM's judgment from the LLM response.
3. **Action Filtering:** If the LLM decision is to not allow the base RL method's action, set $a_t = 0$. Otherwise, set $a_t = \tilde{a}_t$.
4. **Policy Execution and Update:** Take the action a_t. Observe the new partial state $\tilde{s}_{t+1}$ and reward r_t. Update the RL agent's parameters based on the tuple $(\tilde{s}_t, a_t, r_t)$, obtaining θ_{t+1}. Query the participant for an updated state description f_{t+1}.

Applying this framework requires specifying intervention options, specifying the observed state space and reward, selecting a pre-trained LLM model, designing the prompt template, and selecting a base RL method. We will describe the intervention options, state space and rewards that we use in our experiments in the next section. We evaluate several LLMs ranging from 3 to 70 billion parameters. We construct the LLM prompt template for our experiments by combining several components: (B) a description of the adaptive intervention domain and hypothesized behavioral dynamics, (F) the free text participant provided state description, (Q) intermediate reasoning questions to guide LLM inference, and (H) a short trajectory history consisting of the four most recent (state, action, reward)

tuples. Every prompt ends with a final question asking the LLM to make a decision to send the candidate message. We provide an example LLM prompt with all components annotated in Figure 2. We primarily experiment with the BFQH components when trajectory history information is available and the BFQ components otherwise. As noted previously, we use Thompson sampling as the base RL method.

3.2. StepCountJITAI+LLM

We next turn to the design of a simulation environment to evaluate the proposed approach. We extend the base simulator introduced in Karine and Marlin (2024) and described in Section 2 to create an enhanced physical activity messaging-based intervention simulator that generates participant state descriptions using an additional LLM component. We describe each enhancement to the simulator below.

State Augmentation. We introduce a new binary state variable $w_t \in \{0, 1\}$ indicating whether the participant is able to engage in walking or not at time t. We use a Markov chain to simulate w_t. The Markov chain is illustrated in Figure 3 and Table 1. We parameterize the Markov chain via $p_{w_{00}} = P(w_{t+1} = 0|w_t = 0)$, the probability of staying in the "can't walk" state, and $p_{w_{11}} = P(w_{t+1} = 1|w_t = 1)$, the probability of staying in the "can walk" state. All simulations begin with participant in the "can walk" state. The parameters of this Markov chain can then be set to simulate a participant that has a larger or smaller chance of becoming unable to walk, and how long the participant takes on average to recover after they become unable to walk. In our experiments, we consider several different scenarios based on different settings of these parameters.

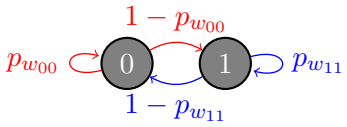

Figure 3: Markov chain sketch.

Table 1: Transition Function.

w_t	w_{t+1}	$P(w_{t+1}\|w_t)$
0	0	$p_{w_{00}}$
0	1	$1 - p_{w_{00}}$
1	0	$1 - p_{w_{11}}$
1	1	$p_{w_{11}}$

Participant Provided State Description Generation. We generate participant provided state descriptions conditioned on the variable w_t using two different LLM prompts. When transitioning from $w_t = 1$ to $w_{t+1} = 0$, we emit text produced by prompting the LLM to generate a short description of a reason why a person might not be able to walk. When staying in the "can't walk" state, no additional state descriptions are generated and the most recent state description stays active. When transitioning from $w_t = 0$ to $w_{t+1} = 1$, we emit text produced by prompting the LLM to generate a message describing that the participant is "feeling fine." When staying in the "can walk" state, we decide whether or not to emit a new participant supplied state description independently at each time point with probability 0.3. This matches average daily survey response rates observed in our past work on year-long physical activity interventions (Spruijt-Metz et al., 2022)

To enable reproducible experiments, we use an LLM to pre-generate lists of simulated participant state descriptions consistent with each state of w_t. Specifically, we generate 500

Example "can walk" state descriptions: I am feeling good, I'm in a great mood, I feel energized, I'm feeling positive, I'm doing well today, I feel great, I'm in high spirits, I feel focused, I'm feeling relaxed, I feel motivated, I'm doing fine, I feel optimistic, I'm feeling calm, I feel balanced, I'm feeling strong, I feel productive, I'm in a positive state of mind, I feel healthy, I feel confident, I feel alert, ...

Example "can't walk" state descriptions: I am tired, I do not want to walk, I got an injury, I have a headache, My legs are sore, I twisted my ankle, I'm feeling dizzy, I'm out of breath, I have a cold, I'm feeling weak, I pulled a muscle, My knee hurts, I have blisters, I feel nauseous, I have stomach cramps, I can't find my shoes, I don't have time, I'm waiting for someone, It's too hot outside, It's too cold outside, ...

Figure 4: Examples of generated participant supplied state descriptions.

state descriptions of each type using ChatGPT (Achiam et al., 2023). Examples of state descriptions generated for each condition are shown in Figure 4.

Behavioral Dynamics. In addition to dynamics for the w_t variable, we need to consider how the "can/can't walk" state should interact with the other state variables already present in the simulator. Below we give the full behavioral dynamics equations for the StepCountJITAI+LLM simulator (new terms are shown in blue) followed by a brief explanation.

$$c_{t+1} \sim Bernoulli(0.5), \quad x_{t+1} \sim \mathcal{N}(c_{t+1}, \sigma^2), \quad p_{t+1} = P(C = 1|x_{t+1}), \quad l_{t+1} = p_{t+1} > 0.5 \tag{3}$$

$$h_{t+1} = \begin{cases} (1 - \delta_h) \cdot h_t & \text{if } a_t = 0 \\ \min(1, h_t + \epsilon_h) & \text{otherwise} \end{cases} \tag{4}$$

$$d_{t+1} = \begin{cases} d_t & \text{if } a_t = 0 \text{ and } w_t = 0 \text{ or } 1 \\ (1 - \delta_d) \cdot d_t & \text{if } a_t \in \{1, c_t + 2\} \text{ and } w_t = 1 \quad \text{(can walk)} \\ \min(1, d_t + \eta_d) & \text{if } a_t \in \{1, c_t + 2\} \text{ and } w_t = 0 \quad \text{(can't walk)} \\ \min(1, d_t + \epsilon_d + (1 - w_t)\ \eta_d) & \text{otherwise} \end{cases} \tag{5}$$

$$z_{t+1} = \begin{cases} m_s + (1 - h_{t+1}) \cdot \rho_1 & \text{if } a_t = 1 \text{ and } w_t = 1 \quad \text{(can walk)} \\ m_s + (1 - h_{t+1}) \cdot \rho_2 & \text{if } a_t = c_t + 2 \text{ and } w_t = 1 \quad \text{(can walk)} \\ m_s\ w_t & \text{otherwise} \end{cases} \tag{6}$$

As in the base simulator, we use c_t to represent a binary context. Habituation increases by additive increments of ϵ_h up to a maximum value of 1 whenever a message is sent and decays by a multiplicative factor of $(1 - \delta_h)$ only when a message is not sent. There is no interaction between the "can/can't walk" state and habituation level. When the participant can walk, the step count z_t depends on the action a_t, the context c_t and the habituation level h_t. When the participant cannot walk, the step count is set to $z_t = 0$.

In the base simulator, the dynamics of disengagement risk are primarily driven by the accuracy of message tailoring (e.g., whether the selected message type matches the current context c_t). The basic intuition is that incorrect tailoring causes the risk of disengagement to increase (for example, due to loss of trust in the intervention system). We similarly

model the effect of sending messages when the participant is in the "can't walk" state as a tailoring error that causes disengagement risk to increase.

The updated dynamics for d_t are a function of the action selected a_t, the context c_t, and walking state w_t. Specifically, if no message is sent to the participant ($a_t = 0$), the disengagement risk remains the same. If a correctly tailored message is sent and the participant can walk, the disengagement risk decreases multiplicatively by a factor of $(1 - \delta_d)$. If a correctly tailored message is sent, but the participant cannot walk, the disengagement risk is incremented by a new parameter η_d up to a maximum value of 1. If an incorrectly tailored message is sent, the disengagement risk is incremented by an amount ϵ_d if the participant can walk, and by an additional amount η_d if the participant cannot walk. As in the base simulator, if the disengagement risk reaches a value of $d_t = 1$, we model the participant as dropping out of the study and the trial ends for that participant.

4. Experiments

In this section, we describe experiments and results. We begin by presenting general experimental protocols. We then describe individual experiments and discuss results.

4.1. Experimental Protocols

All experiments use the StepCountJITAI+LLM simulator. In all experiments, we set $\delta_h = 0.1$, $\epsilon_h = 0.05$, $\delta_d = 0.1$, $\epsilon_d = 0.05$, $\rho_1 = 50$, $\rho_2 = 200$. This yields a modest increase in steps for providing correctly tailored intervention content. We give results in terms of excess steps above the baseline step count m_s. We vary the $p_{w_{11}}$, $p_{w_{00}}$, and η_d parameters to simulate participants with different characteristics. The full state in all experiments corresponds to $s_t = [c_t, h_t, d_t, w_t]$. The observed state accessible to the base RL agent in all experiments is $\tilde{s}_t = [c_t, h_t, d_t]$. Thus, w_t is latent from the perspective of the base RL agent. We simulate a trial of up to 50 days with one decision point per day. The reward is set to 0 for any time steps following a disengagement event. For both LLM4TS and TS, all learning occurs within-trial for an individual participant. There is no sharing of data between participants. We set the TS prior parameters to $\mu_{0a} = 0$ and $\Sigma_{0a} = 100I$ for each action a and the reward noise variance $\sigma^2_{Ya} = 25^2$ for each action a (see Equations 1 and 2).

The participant provided state descriptions f_t are a function of w_t only as described earlier. We pre-generate a total of 1000 state descriptions using ChatGPT, 500 for each walking ability state. We sample from this set when running the simulator. For experiments using LLM inference, we evaluate Llama 3 70B, Llama 3 8B and Gemma 2 9B (Llama Team, 2024; Gemma Team, 2024). We use the prompt components shown in Figure 2 for all experiments. All learning experiments are repeated five times and results are given in terms of median performance along with the 25th and 75th percentiles.

4.2. Validating LLM Generation

The first question we pose is, are the ChatGPT generated descriptions consistent with the intended state values? To answer this question, we extract a data set at random from the 1000 state descriptions containing 100 total state descriptions including 50 generated using the "cannot walk" prompt and 50 generated using the "feeling fine" prompt. We blinded

Table 2: Inference results using different LLMs.

Model	Accuracy	Precision	Recall	F1
Llama 3 70B (BFQ)	0.999	0.998	0.999	0.999
Llama 3 8B (BFQ)	0.881	0.992	0.881	0.866
Gemma 2 9B (BFQ)	0.918	0.995	0.918	0.911

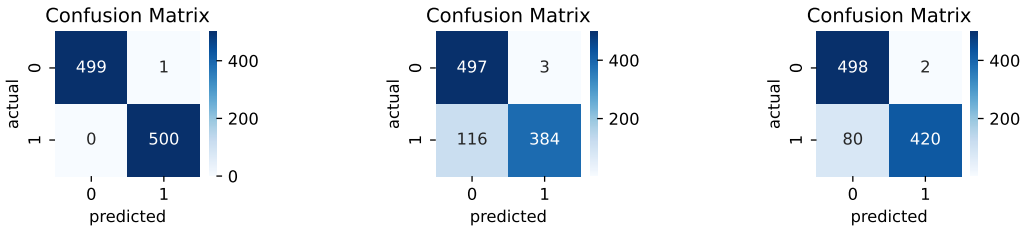

Figure 5: Confusion matrices (left to right): Llama 3 70B, Llama 3 8B, Gemma 2 9B.

the generating labels and manually classified each state description. This evaluation showed that 100% of the generated state descriptions were consistent with the intended state. This verifies that ChatGPT can generate descriptions consistent with the desired states.

4.3. Validating LLM Inference

The next question we ask is how well do different LLMs perform at the task of inferring whether messages should be sent based on simulated participant provided state descriptions using the developed prompt. We consider the true label to be "send" when the walking state is "can walk" and the true label to be "don't send" when the walking state is "can't walk." We evaluate Llama 3 70B, Llama 3 8B and Gemma 2 9B (Llama Team, 2024; Gemma Team, 2024). We use an LLM temperature of 0.2 (level of randomness in the LLM response) and the BFQ prompt strategy. We give confusion matrices for this inference problem in Figure 5. We summarize accuracy, precision, recall and F1 metrics in Table 2. We can see that while accuracy is positively correlated with model size among the three models tested, all three LLMs achieve accuracy above 85%. For Llama 3 8B and Gemma 2 9B, we can see that the drop in accuracy is mostly accounted for by lower recall where these models incorrectly infer that messages should not be sent when sending messages is actually allowable.

4.4. Evaluating LLM4TS

We conduct extensive experiments to compare LLM4TS to standard Thompson Sampling (TS). We explore four different scenarios for the dynamics of w_t. Scenario 1: $p_{w_{11}} = 0.7$, $p_{w_{00}} = 0.5$. Scenario 2: $p_{w_{11}} = 0.7$, $p_{w_{00}} = 0.1$, Scenario 3: $p_{w_{11}} = 0.95$, $p_{w_{00}} = 0.5$. Scenario 4: $p_{w_{11}} = 0.95$, $p_{w_{00}} = 0.1$. Among these scenarios, Scenario 1 has the highest per-time step chance of the participant becoming unable to walk, as well as the lowest chance of recovering. Scenario 4 has the lowest per-time step chance of the participant becoming

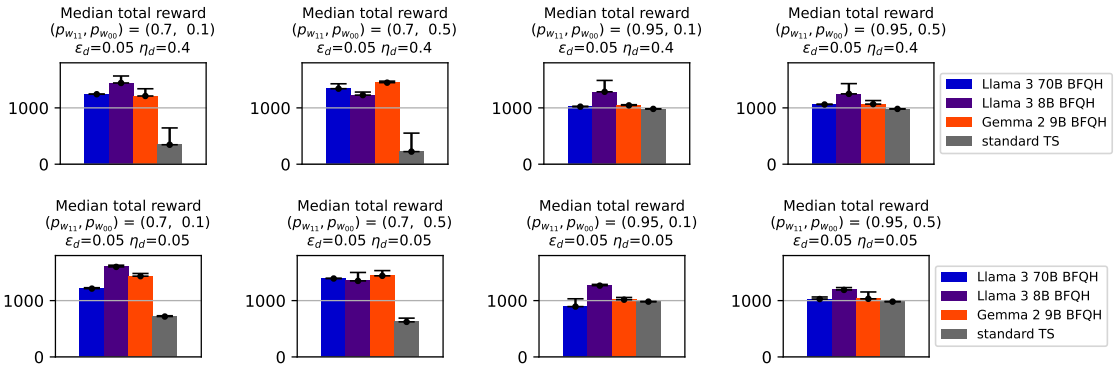

Figure 6: Comparing LLMs. Columns correspond to different choices for $(p_{w_{11}}, p_{w_{00}})$. Rows correspond to different choices for η_d. Bar colors indicate different approaches.

unable to walk, as well as the highest chance of recovering. We also experiment with the value of η_d, the parameter controlling the disengagement risk increment when messages are sent in the "can't walk" state. We consider the setting $\eta_d = 0.05$, which corresponds to a participant whose disengagement risk increases very mildly in response to receiving messages in the "can't walk" state, and $\eta_d = 0.4$, which corresponds to a participant whose disengagement risk increases significantly.

Comparing LLM4TS and standard TS. The question of interest in this experiment is under what circumstances can the LLM4TS method improve on standard TS? We compare standard TS to LLM4TS using Llama3 70B, Llama3 8B and Gemma2 9B and the BFQH prompt. We show the results in Figure 6. Each row corresponds to a disengagement risk increment parameter value η_d. Each column corresponds to a different scenario for the w_t dynamics. The bars correspond to different methods. The first trend we can see in these results is that when $p_{w_{11}} = 0.7$, indicating a higher probability of transitioning to the can't walk state, LLM4TS exhibits significantly higher average reward than standard TS. This suggests that the LLM inference step and action filter are indeed contributing to improving adaptive intervention performance when participants are in the "can't walk" state.

Second, with $p_{w_{11}} = 0.7$ we can observe a more subtle trend: the performance difference between LLM4TS and standard TS increases as the disengagement risk increment parameter η_d is increased. This again makes sense as standard TS will trigger the disengagement risk threshold more often the higher the value of this parameter and therefore should obtain lower reward. The performance of LLM4TS does not vary significantly with η_d since it is largely able to avoid sending messages in the "can't walk" state.

Third, when $p_{w_{11}} = 0.95$ indicating a low chance of transitioning to the "can't walk" state, the performance of LLM4TS and standard TS are much closer regardless of the values of η_d and $p_{w_{00}}$. This should be expected as LLM4TS has limited opportunities to improve performance when participants rarely enter the "can't walk" state.

Lastly, we observe that when $p_{w_{11}} = 0.7$, the total reward for LLM4TS is higher than when $p_{w_{11}} = 0.95$. This is counter intuitive since the reward is zero when the participant can't walk. However, inferring that a messages should not be sent results in the habituation level decreasing. Since the base TS agent selects messages to optimize expected immediate

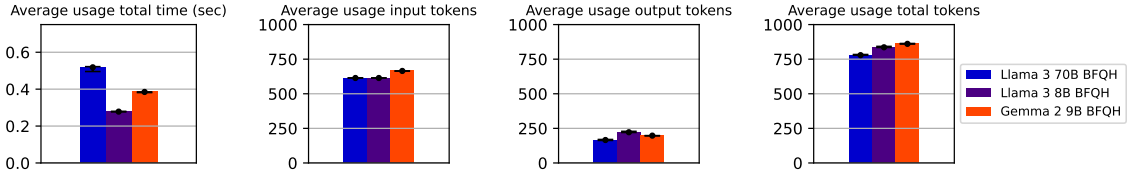

Figure 7: Comparing LLM metrics for $\eta_d = 0.4$ and $(p_{w_{11}}, p_{w_{00}}) = (0.7, 0.1)$.

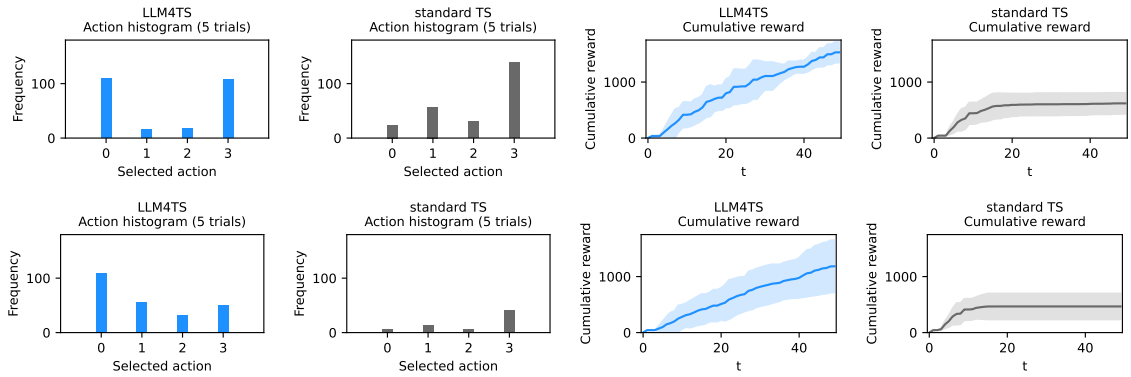

Figure 8: Histograms of action values and plots of average cumulative reward per episode for $(p_{w_{11}}, p_{w_{00}}) = (0.7, 0.1)$, with prompt strategy BFQH. (Top row) $\eta_d = 0.05$. (Bottom row) $\eta_d = 0.4$. (Blue) Llama 3 8B. (Gray) standard TS.

reward and the true expected immediate reward of sending a message is always higher than not sending a message, the base TS agent tends to send more messages than is optimal when taking into account the long term effect of actions. The LLM4TS action filter thus has a secondary positive effect due to decreasing habituation when standard TS is used as the base RL agent. Interestingly, this observation may also explain why Llama 3 70B has lower end-to-end performance despite having the best inference accuracy in the previous experiment.

LLM Runtime Metrics. In this assessment, we examine the question of how runtime metrics differ for the LLMs used in the previous experiment. We examine average per-call LLM inference time, average per-call input token count, and average per-call total (input+output) token count. The results are shown in Figure 7. We can see that the number of input tokens is similar for all LLMs as expected. On the other hand, the average time per LLM inference call using LLama 3 8B is much lower than when using the other two LLMs, while the average number of output tokens is only modestly higher than Llama 3 70B. However, the per-output token cost for LLama 3 8B is currently much lower than for LLama 3 70B on commercial LLM API providers, offsetting the higher output token count. These results combined with the results of the previous experiment indicate that LLama 3 8B offers strong performance in terms of average reward, inference time and cost. In the following experiments, we focus on Llama 3 8B.

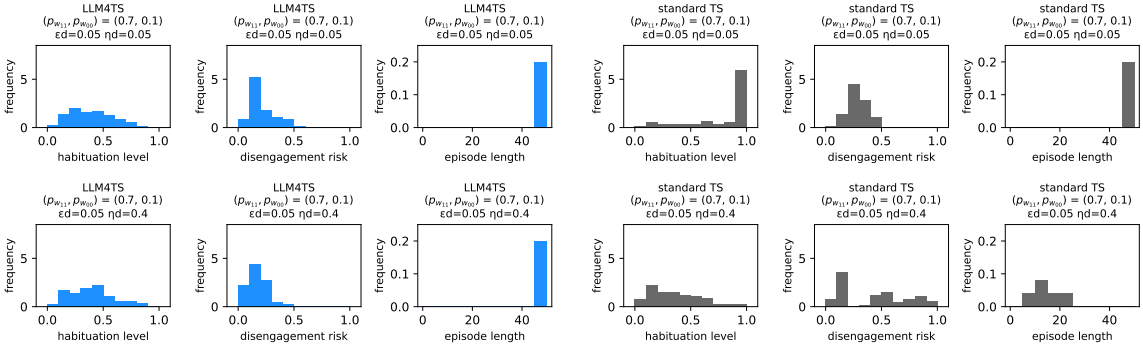

Figure 9: Histograms of habituation level, disengagement risk, and episode length for $(p_{w_{11}}, p_{w_{00}}) = (0.7, 0.1)$, with prompt strategy BFQH. (Top row) $\eta_d = 0.05$. (Bottom row) $\eta_d = 0.4$. (Blue) Llama 3 8B. (Gray) standard TS.

Analysis of Actions and States. In this section, the primary question we address is the mechanism by which LLM4TS outperforms TS in different scenarios. We first analyze the actions selected by LLM4TS compared to TS. We focus on $p_{w_{11}} = 0.7$ and $p_{w_{00}} = 0.1$ as a case where LLM4TS and TS differ in performance. We consider $\eta_d = 0.05$ and $\eta_d = 0.04$. We show results for Llama 3 8B in Figure 8. The action selection histograms show that LLM4TS selects more $a_t = 0$ actions, which indicates that LLM4TS has decided to not send a message more often than standard TS. We also compare the average cumulative reward per episode. The plateau seen in the TS cumulative reward plot suggests that TS is incurring disengagement events.

Next, we analyze the distribution of habituation level, disengagement risk and episode length. We first note that for $\eta_d = 0.05$, all trials for both methods complete all 50 time steps. This is possible for TS due to the low penalty on disengagement risk when sending messages while the participant cannot walk. However, we can see that the distribution of habituation values is concentrated on much higher values for TS than for LLM4TS, which results from sending messages during period where the participant cannot walk. This explains how LLM4TS outperforms TS in terms of total reward in this scenario. Next, when $\eta_d = 0.4$, we can see that the distribution of habituation for plain TS is actually lower than for LLM4TS. However, the disengagement risk values are much higher and indeed we can see that none of the TS trials reach the full episode length indicating that all trials hit the disengagement risk threshold. Again, this can be attributed to TS sending messages when the participant cannot walk, while LLM4TS appropriately filters these actions.

Comparing Prompt Structures. Lastly, we perform a comparison of different prompt structures. The primary question of interest is how does removing components from the BFQH prompt structure affect total reward? We consider the alternative prompt structures BFQ, and BF. The BFQH prompt structure corresponds to the example shown in Figure 2. The BFQ prompt structure removes the trajectory history. The BF prompt structure further removes the intermediate reasoning questions. We show the results for Llama 3 8B in Figure 10. The columns again correspond to different choices for $p_{w_{11}}$, and $p_{w_{00}}$. We show results for $\eta_d = 0.4$ and $\eta_d = 0.05$. The results show that the BFQH prompt structure has

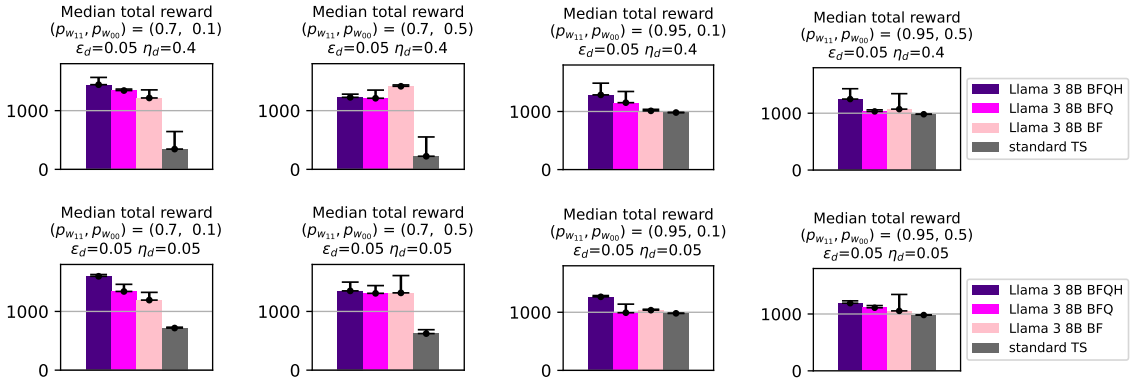

Figure 10: Comparing LLM prompt strategies, for various (p_{w11}, p_{w00}), using Llama 3 8B, with $\eta_d = 0.4$ (top), $\eta_d = 0.05$ (bottom). Standard TS is shown for comparison.

median performance that is better than the next best prompt structure in six of the eight scenarios and is only outperformed in one scenario. These results suggest that performance degrades on average when removing components from the BFQH prompt structure.

5. Discussion

In this work, we have presented LLM4TS, an approach to augmenting a base RL method with LLM-based reasoning capabilities to help address the significant challenge of data scarcity that arises when applying RL methods to optimize adaptive intervention policies in the context of practical research study designs. Further, we have developed a physical activity JITAI simulation environment, StepCountJITAI+LLM, that models key behavioral dynamics as well as the generation of participant-provided state descriptions via an auxiliary LLM. We have presented experiments and results validating the generation of participant-provided state descriptions, the ability of LLMs to infer states from state descriptions, and the clear benefits of LLM4TS over standard Thompson sampling in scenarios where the LLM reasoning component is exercised. These results support our claim that the LLM4TS approach is able to improve the limited state representation of a base Thompson sampler while maintaining data efficiency.

We emphasize that the LLM4TS approach is a general and broadly applicable framework for enhancing the intelligence of adaptive interventions. The approach can be applied to different adaptive intervention domains by engineering appropriate LLM inference prompts. Our results suggest that supplying trajectory histories and intermediate reasoning questions along with participant provided state descriptions and hypotheses about behavioral dynamics contribute to improving performance. There is wide leeway for modifications such as supplying additional domain specific knowledge and investigating alternative intermediate reasoning questions. The approach can also be combined with any instruction tuned LLM and can thus benefit from future advances in LLM models. Similarly, while we have focused on applying this framework with a Thompson sampler as the base RL method, it can be applied with any data-efficient RL method.

Limitations. While we believe our results show that the proposed approach holds significant promise for augmenting the intelligence of adaptive health interventions, this study has several limitations. First, our evaluation is limited to exploring the performance of the proposed approach in the context of a physical activity adaptive intervention simulation. While we expect LLMs to have sufficient world knowledge to provide appropriate reasoning in other behavioral intervention domains, this requires validation. For intervention domains where LLMs lack sufficient prior world knowledge, such knowledge could be provided as part of a reasoning prompt. Second, the utility of the approach hinges on the willingness of participants to provide state information to the intervention system. Our simulations do not assume that participants always provide responses when prompted, but real applications may need to contend with additional issues like informative missingness where participants are unlikely to respond when highly significant life events occur. Lastly, the magnitude of the performance improvements we observe in our experiments depends on many details of the simulation environment and we have highlighted multiple scenarios showing different levels of performance improvement. While validating this approach in simulation is an important first step, the next step for the approach requires evaluation in a human subjects study. The evidence presented in this paper will establishes the foundation for conducting such a study.

Acknowledgments

This work was supported by the National Institute of Biomedical Imaging and Bioengineering and the National Institute on Aging through grants 1P41EB028242 and P30AG073107. The authors thank Dr. Susan Murphy, Dr. Pedja Klasnja, and Dr. Steven De La Torre for helpful discussions of this work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

Linda M Collins, Susan A Murphy, and Victor Strecher. The multiphase optimization strategy (most) and the sequential multiple assignment randomized trial (smart): new methods for more potent ehealth interventions. *American journal of preventive medicine*, 32(5):S112–S118, 2007.

Antonio Coronato, Muddasar Naeem, Giuseppe De Pietro, and Giovanni Paragliola. Reinforcement learning for intelligent healthcare applications: A survey. *Artificial Intelligence in Medicine*, 109:101964, 2020.

Milan Radovan Dimitrijević, Janez Faganel, Matej Gregorić, PW Nathan, and JK Trontelj. Habituation: effects of regular and stochastic stimulation. *Journal of Neurology, Neurosurgery & Psychiatry*, 35(2):234–242, 1972.

Mohamed El Mistiri, Owais Khan, César A. Martin, Eric Hekler, and Daniel E. Rivera. Data-driven mobile health: System identification and hybrid model predictive control to deliver personalized physical activity interventions. *IEEE Open Journal of Control Systems*, 4:83–102, 2025.

Evan M Forman, Stephanie P Goldstein, Rebecca J Crochiere, Meghan L Butryn, Adrienne S Juarascio, Fengqing Zhang, and Gary D Foster. Randomized controlled trial of ontrack, a just-in-time adaptive intervention designed to enhance weight loss. *Translational behavioral medicine*, 9(6):989–1001, 2019.

Gemma Team. Gemma: Open Models Based on Gemini Research and Technology. *arXiv:2403.08295*, 2024.

Stephanie P Goldstein, Fengqing Zhang, Predrag Klasnja, Adam Hoover, Rena R Wing, and John Graham Thomas. Optimizing a just-in-time adaptive intervention to improve dietary adherence in behavioral obesity treatment: protocol for a microrandomized trial. *JMIR research protocols*, 10(12):e33568, 2021.

Farid Golnaraghi and Benjamin C Kuo. Automatic control systems. *Complex Variables*, 2: 1–1, 2010.

Suat Gönül, Tuncay Namlı, Ahmet Coşar, and İsmail Hakkı Toroslu. A reinforcement learning based algorithm for personalization of digital, just-in-time, adaptive interventions. *Artificial Intelligence in Medicine*, 115:102062, 2021.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Wendy Hardeman, Julie Houghton, Kathleen Lane, Andy Jones, and Felix Naughton. A systematic review of just-in-time adaptive interventions (jitais) to promote physical activity. *International Journal of Behavioral Nutrition and Physical Activity*, 16(1):1–21, 2019.

Halsted R Holman. The relation of the chronic disease epidemic to the health care crisis. *ACR open rheumatology*, 2(3):167–173, 2020.

Hengyuan Hu and Dorsa Sadigh. Language instructed reinforcement learning for human-ai coordination. In *International Conference on Machine Learning*, pages 13584–13598. PMLR, 2023.

Karine Karine and Benjamin M. Marlin. StepCountJITAI: simulation environment for RL with application to physical activity adaptive intervention. In *Workshop on Behavioral Machine Learning, Advances in Neural Information Processing Systems*, 2024.

Karine Karine, Predrag Klasnja, Susan A. Murphy, and Benjamin M. Marlin. Assessing the impact of context inference error and partial observability on RL methods for Just-In-Time Adaptive Interventions. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216, pages 1047–1057, 2023.

Taewoo Kim, Minsu Jang, and Jaehong Kim. A survey on simulation environments for reinforcement learning. In *2021 18th International Conference on Ubiquitous Robots (UR)*, pages 63–67, 2021. doi: 10.1109/UR52253.2021.9494694.

Korbinian Kuhn, Verena Kersken, Benedikt Reuter, Niklas Egger, and Gottfried Zimmermann. Measuring the accuracy of automatic speech recognition solutions. *ACM Transactions on Accessible Computing*, 16(4):1–23, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.

Peng Liao, Walter Dempsey, Hillol Sarker, Syed Monowar Hossain, Mustafa al'Absi, Predrag Klasnja, and Susan Murphy. Just-in-time but not too much: Determining treatment timing in mobile health. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(4), December 2018.

Peng Liao, Kristjan Greenewald, Predrag Klasnja, and Susan Murphy. Personalized HeartSteps: A Reinforcement Learning Algorithm for Optimizing Physical Activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1): 1–22, 2020.

Llama Team. The Llama 3 Herd of Models. *arXiv:2407.21783*, 2024.

Inbal Nahum-Shani, Shawna N Smith, Bonnie J Spring, Linda M Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A Murphy. Just-in-time adaptive interventions (jitais) in mobile health: key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine*, 52(6):446–462, 2018.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022.

Joonyoung Park and Uichin Lee. Understanding disengagement in just-in-time mobile health interventions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(2):1–27, 2023.

Olga Perski, Emily T Hébert, Felix Naughton, Eric B Hekler, Jamie Brown, and Michael S Businelle. Technology-mediated just-in-time adaptive interventions (JITAIs) to reduce harmful substance use: a systematic review. *Addiction*, 117(5):1220–1241, 2022.

Moschoula Pternea, Prerna Singh, Abir Chakraborty, Yagna Oruganti, Mirco Milletari, Sayli Bapat, and Kebei Jiang. The rl/llm taxonomy tree: Reviewing synergies between reinforcement learning and large language models. *Journal of Artificial Intelligence Research*, 80, 2024.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

Daniel J. Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A Tutorial on Thompson Sampling. *Found. Trends Mach. Learn.*, 11(1), 2018.

Donna Spruijt-Metz, Benjamin M Marlin, Misha Pavel, Daniel E Rivera, Eric Hekler, Steven De La Torre, Mohamed El Mistiri, Natalie M Golaszweski, Cynthia Li, Rebecca Braga De Braganca, et al. Advancing behavioral intervention and theory development for mobile health: the HeartSteps II protocol. *International journal of environmental research and public health*, 19(4):2267, 2022.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

William R. Thompson. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. In *Biometrika*, volume 25, pages 285–294, 1933.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.

Min-Jeong Yang, Steven K Sutton, Laura M Hernandez, Sarah R Jones, David W Wetter, Santosh Kumar, and Christine Vinci. A just-in-time adaptive intervention (jitai) for smoking cessation: Feasibility and acceptability findings. *Addictive Behaviors*, 136: 107467, 2023.

Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623, 2023.

Appendix A. StepCountJITAI simulation environment

The base simulator introduced in Karine et al. (2023); Karine and Marlin (2024) models the behavior of a participant in a messaging-based mobile health study. The intervention options (actions) are the messages sent to the participant, with the goal of increasing the participant's walking step count (reward), given the participant's context (state). We summarize the base simulator specification in Tables 3 and 4, and provide details below.

A.1. StepCountJITAI specifications

In this sections, we use uppercase letters for variable names, and lowercase letters for variable values. c_t denotes the true context, p_t is the probability of context 1, l_t is the inferred context, h_t is the habituation level, d_t is the disengagement risk, z_t is the participant's walking step count, and a_t is the action at time t. The base simulator also includes behavioral parameters: δ_d and ϵ_d are decay and increment parameters for the disengagement risk, and δ_h and ϵ_h are decay and increment parameters for the habituation level. The goal of the simulated intervention is to maximize the total walking step count. Thus, the walking step count is also the RL reward ($r_t = z_t$).

Table 3: Environment state variables

Variable	Description	Values
c_t	True context	$\{0, 1\}$
p_t	Probability of context 1	$[0, 1]$
l_t	Inferred context	{0,1}
d_t	Disengagement risk level	$[0, 1]$
h_t	Habituation level	$[0, 1]$
z_t	Walking step count	$\mathbb{N}$

Table 4: Possible action values

Action	Description
$a_t = 0$	No message is sent to the participant.
$a_t = 1$	A non-contextualized message is sent.
$a_t = 2$	A message customized to context 0 is sent.
$a_t = 3$	A message customized to context 1 is sent.

A.2. StepCountJITAI behavioral dynamics

The behavioral dynamics of the base simulator are as follow: Sending a message causes the habituation level to increase. Not sending a message causes the habituation level to decrease. An incorrectly tailored message causes the disengagement risk to increase. A correctly tailored message causes the disengagement risk to decrease. When the disengagement risk exceeds a given threshold, the behavioral study ends. The reward is the surplus walking

step count, beyond a baseline count, attenuated by the habituation level. The behavioral dynamics equations for the base simulator are provided below. σ is the context uncertainty, x_t is a context feature. $\sigma, \rho_1, \rho_2, m_s$ are fixed parameters.

$$c_{t+1} \sim \mathit{Bernoulli}(0.5), \quad x_{t+1} \sim \mathcal{N}(c_{t+1}, \sigma^2), \quad p_{t+1} = P(C = 1 | x_{t+1}), \quad l_{t+1} = p_{t+1} > 0.5 \tag{7}$$

$$h_{t+1} = \begin{cases} (1 - \delta_h) \cdot h_t & \text{if } a_t = 0 \\ \min(1, h_t + \epsilon_h) & \text{otherwise} \end{cases} \tag{8}$$

$$d_{t+1} = \begin{cases} d_t & \text{if } a_t = 0 \\ (1 - \delta_d) \cdot d_t & \text{if } a_t \in \{1, c_t + 2\} \\ \min(1, d_t + \epsilon_d) & \text{otherwise} \end{cases} \tag{9}$$

$$z_{t+1} = \begin{cases} m_s + (1 - h_{t+1}) \cdot \rho_1 & \text{if } a_t = 1 \\ m_s + (1 - h_{t+1}) \cdot \rho_2 & \text{if } a_t = c_t + 2 \\ m_s & \text{otherwise} \end{cases} \tag{10}$$