Sleepless Nights, Sugary Days: Creating Synthetic Users with Health Conditions for Realistic Coaching Agent Interactions

Taedong Yun[1,†], Eric Yang[2], Mustafa Safdari[1], Jong Ha Lee[2], Vaishnavi Vinod Kumar[3], S. Sara Mahdavi[1], Jonathan Amar[2], Derek Peyton[3], Reut Aharony[1], Andreas Michaelides[3], Logan Schneider[3], Isaac Galatzer-Levy[3], Yugang Jia[2], John Canny[1], Arthur Gretton[1,*,†], Maja Matarić[1,*,†]

[1]Google DeepMind [2]Verily Life Sciences [3]Google

[*]Co-last authors (listed alphabetically)

[†]{tedyun,gretton,majamataric}@google.com

Abstract

We present an end-to-end framework for generating synthetic users for evaluating interactive agents designed to encourage positive behavior changes, such as in health and lifestyle coaching. The synthetic users are grounded in health and lifestyle conditions, specifically sleep and diabetes management in this study, to ensure realistic interactions with the health coaching agent. Synthetic users are created in two stages: first, structured data are generated grounded in real-world health and lifestyle factors in addition to basic demographics and behavioral attributes; second, full profiles of the synthetic users are developed conditioned on the structured data. Interactions between synthetic users and the coaching agent are simulated using generative agent-based models such as Concordia, or directly by prompting a language model. Using two independently-developed agents for sleep and diabetes coaching as case studies, the validity of this framework is demonstrated by analyzing the coaching agent's understanding of the synthetic users' needs and challenges. Finally, through multiple blinded evaluations of user-coach interactions by human experts, we demonstrate that our synthetic users with health and behavioral attributes more accurately portray real human users with the same attributes, compared to generic synthetic users not grounded in such attributes. The proposed framework lays the foundation for efficient development of conversational agents through extensive, realistic, and grounded simulated interactions.

1 Introduction

Personalization is a key capability for any interactive agents aiming to encourage positive behavior changes, for instance in health and lifestyle (Christakopoulou et al., 2024; Yang et al., 2024). The efficacy of such agents must be assessed via interactions with users; however, collecting and evaluating diverse, long-term human interactions with agents is costly and time-consuming, and the development of such agents can be significantly accelerated with realistic simulation. We address the challenge of realistic user simulation in this context by proposing an end-to-end framework for generating cohorts of synthetic users grounded in health, lifestyle, and behavior, and having them interact with coaching agents. The framework is generally applicable across a range of health and lifestyle domains; we validate it in the domains of sleep and diabetes coaching.

Synthetic users generated with large language models (LLMs) have been widely studied in recent years (Kapania et al., 2024; Wang et al., 2025; Moon et al., 2024; Park et al., 2024, 2023), often for general purpose alignment and personalization (Castricato et al., 2024), and have been designed to reflect demographic information about the population at large. We focus on a more targeted application, where the synthetic users are employed for interaction with foundation models or agents specialized in health, as done in recent work (Tu et al., 2024; Yu et al., 2024; Johri et al., 2025). We provide an overview of such work in Section 2.

We demonstrate an end-to-end-framework for designing a cohort of synthetic users exhibiting health conditions consistent with their specific demographics and behavioral factors, in order for their interaction with a coaching agent to realistically reflect their needs and challenges. We construct the synthetic users by first generating a sample of natural language "vignettes", which are in turn generated based on real demographics, health, and behavioral data (such as the Big Five markers; Goldberg, 1992; Serapio-García et al., 2023). Our framework also allows for optionally including additional information: domain-specific rubrics

Findings of the Association for Computational Linguistics: ACL 2025, pages 14159–14181

such as “barriers” and “goals” as used by human coaches; challenges to action derived from the COM-B behavioural model (Michie et al., 2011); or rich backstories for the synthetic users, generated consistently with the structured attributes from real data. Finally, based on this cohort of vignettes, realistic user-coach simulation can be performed by direct LLM calls or by using generative agent-based models such as the Concordia (Vezhnevets et al., 2023) open-source system. More details will be discussed in Section 3.

We demonstrate the efficacy of our methodology for generating health-grounded synthetic personas in two separate domains, sleep coaching (Section 4) and diabetes coaching (Section 5), using two independently developed sleep coaching and diabetes coaching agents. We verify that our synthetic users indeed communicate behaviors, health conditions, and barriers consistent with their assigned attributes, by inspecting the internal state-of-user model of the coaching agent for each synthetic user, by a holistic evaluation of the user-coach interactions by trained human experts, and by a comparative evaluation of our health-grounded synthetic users against generic synthetic users, again by trained human experts.

2 Background

In this section, we briefly survey the development and goals of LLM-based synthetic personas in general, and then provide further details on prior literature on the design and application of synthetic personas for health. We next describe software systems for simulating agent interactions. Finally, we discuss some challenges in using LLM-derived synthetic users for agent development and beyond.

2.1 LLMs as Synthetic Personas

The time and expense required in human evaluation of LLM agents naturally evokes the need for auto-evaluation. This has led to the development of “reinforcement learning from AI feedback (RLAIF)”, where LLMs serve to evaluate the output of other LLMs, in a now well-established strategy for maximizing alignment, factuality, helpfulness, and other desiderata (Bai et al., 2022; Lee et al., 2024).

One recent approach to RLAIF uses LLMs to construct synthetic personas and to judge their interaction with the model being evaluated. Castricato et al. (2024) generated these synthetic personas from the US census data, and employed them to build an evaluation dataset of prompt and feedback pairs obtained solely from synthetic users. A strength of the approach is that results for specific users can be modeled, rather than general user categories. This approach also motivates the work of Moon et al. (2024).

In addition to demographic information, Serapio-García et al. (2023) showed that LLMs can be designed to display specific personality traits, such as those along the Big Five dimensions (Goldberg, 1992). Effective modeling and control of personality in LLMs is important for two reasons: first since personality is a key factor in determining effective communication; second, because it is important to model interaction of synthetic users over the broad range of personality profiles to be encountered in practice. Serapio-García et al. (2023) found that instruction-tuned models displayed more reliable and externally valid personality types than pre-trained variants (perhaps due to superior instruction following), and that larger models are better equipped to express complex traits, and to simulate social behaviours.

2.2 LLMs and Synthetic Personas in Health

Several works have proposed the use of LLMs to provide advice in health or wellness settings.

Tu et al. (2024) introduced Articulate Medical Intelligence Explorer (AMIE), an LLM agent designed for conversational medical diagnosis. The AMIE agent achieved remarkable success, with greater diagnostic accuracy compared to primary care physicians in a randomized, double-blind crossover study. AMIE was trained using self-play with simulated doctor-patient dialogues and was tested with patients simulated first by searching for medical conditions from across three databases (Health QA, MalaCards Human Disease, MedicineNet Diseases & Conditions), and then by performing internet searches on each medical condition in order to retrieve passages on demographics, symptoms, and management, which were screened for relevance. PaLM 2 (Google, 2023) was used as the base model for all agents, including the patients and doctors, which were given role-specific prompts. Instruction fine tuning of both patient and doctor took place, first using static datasets of real doctor-patient interactions, and subsequently using dialogues generated through self-play. A moderator agent is used to manage the dialogue between doctor and patient, and to decide on when the conversation has ended.

There were a number of limitations to the synthetic user (patient) design in the AMIE study. First, the retrieval of symptoms and demographics from online discussion might create bias towards demographics of individuals most likely to engage in online discussion of their condition. Second, online discussion might misattribute irrelevant symptoms to conditions, and be misinformed as to mitigation strategies. Third, the model did not explicitly control for personality traits, relying on the distribution of personality traits to be found in online discussion of the medical conditions, which might not be representative of target demographics. Related to this issue, PaLM 2 has undergone a fine-tuning process which may make the patient agents more obliging, sympathetic, and communicative than real patients.

Yu et al. (2024) also described an application of LLMs to simulate user (patient) populations, based on particular medical histories from 1495 patients in the MIMIC-III database (Johnson et al., 2016). The approach represented each patient's medical and demographic status as a knowledge graph (with 15,000 possible nodes and over 26,000 possible edges), then retrieved patient information from this graph in response to doctor queries. The retrieved answers were then expressed in natural language, with tone adapted in accordance with the Big Five personality traits. The key finding was that "patient LLMs" that retrieve from a structured representation of patient traits more accurately communicate symptoms and history than simply using unstructured clinical notes in the prompt.

This past work differs from our approach in a number of important ways. First, in the past work, the representation of each individual as a knowledge graph is suited to clinical settings, where unstructured clinical notes must be organized as a preprocessing step. Accordingly, much of the focus of their work is on the construction of the graph, and accurate retrieval from it. The graph representation may be less suited to wellness settings, however, where the user's needs and challenges require less structure and granularity. Second, and more importantly, the generation process can only traverse the graph of existing patients, and cannot generate new individuals. This may lead to limitations with respect to both privacy and scalability.

Johri et al. (2025) described the use of synthetic personas in evaluating an LLM system for patient history collection in medicine. The patient agent was instructed explicitly not to display medical knowledge or generate new symptoms beyond what was given in the "vignette". A total of 2,000 case vignettes were considered: 1,800 questions from MedQA-United States Medical Licensing Examination covering 12 medical specialties, and 200 questions on skin disease from the Derm-Public and Derm-Private question banks. While representing a broad range of conditions, the focus of that study was primarily on evaluating information gathering and diagnosis abilities of LLMs, rather than coaching. Unlike our work, Johri et al. did not attempt to ensure that patient profiles were representative of the population at large, nor were personality types modeled when generating patient dialogues.

2.3 Health Coaching Datasets

There are a limited number of datasets on patient conversations with health practitioners, such as Zhou et al. (2024) and Wang et al. (2023).

In brief, the Zhou et al. (2024) dataset consists of a cohort of 22 patients for up to eight weeks, and records both the initial S.M.A.R.T. (specific, measurable, achievable, relevant, and time-bound) goal creation, and subsequent progress monitoring and engagement. Zhou et al. (2024) focus on a neurosymbolic representation of goals negotiated in collaboration with a coach, which can include very broad wellness categories. On the other hand, our focus is on more sharply defined wellness issues (sleep or diet) where specific goals and challenges are addressed. We also address only the elicitation of barriers and goals, not subsequent behaviour change.

Wang et al. (2023) provides a dataset of 68 patient dialogues in the setting of engagement with medical practitioners for managing cancer symptoms. Their use of LLMs to monitor engagement is a promising direction for future updates to the coaching agent in our study.

2.4 Generative Agent-Based Models

Generative agent-based models, such as the Concordia system which we used (Vezhnevets et al., 2023), are directly relevant to user persona simulation. Concordia offers a framework for generating agent interactions, particularly dialog-based ones, given a sufficient backstory. Its design incorporates several key features beneficial for research in generative agent-based simulation. First, Concordia employs associative memory (Park et al., 2023) and chain-of-thought reasoning (Wei et al., 2022) to ensure that agent utterances are grounded in the provided

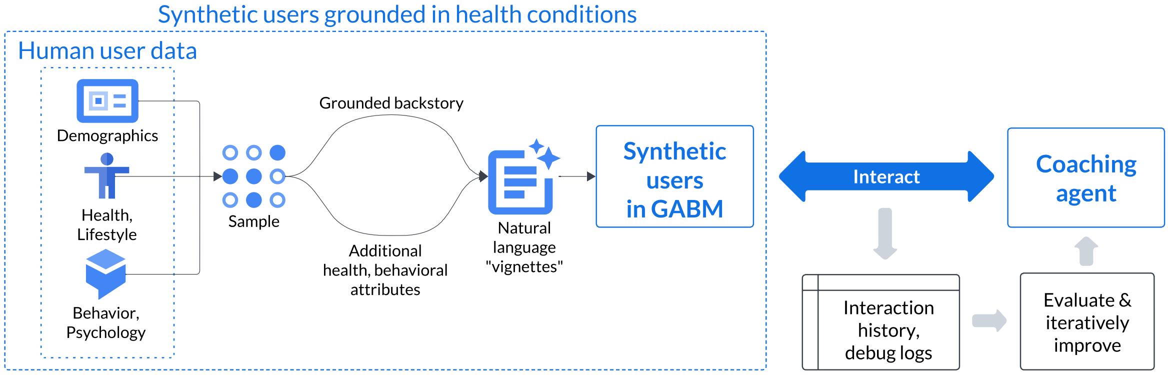

Figure 1: An overview of generating synthetic users grounded in real demographics, health & lifestyle, and behavioral & psychological characteristics for automated evaluation of coaching agent interactions. GABM: generative agent-based model.

backstory. Second, it utilizes metadata tags within prompts to clearly delineate different contextual elements, including agent memory, observations, prior conversation turns, and overarching goals. Furthermore, these contextual components can be assigned varying levels of importance, guiding the underlying LLM to generate subsequent dialog using appropriate chain-of-thought reasoning. Third, Concordia provides extensive logging capabilities, enabling detailed examination of the pipeline's internal processes, such as the constructed prompts sent to the LLM, the chain-of-thought reasoning process, post-processing steps, and the final output. Fourth, its modular architecture allows for flexible configuration, enabling researchers to customize the agent pipeline by swapping components.

A crucial requirement for generative agent-based simulation frameworks is the ability to integrate and manage multiple, mutable states within the interaction environment, as these states can significantly influence an agent's subsequent actions and utterances. Critically, the mutation of these states, often lacking a true real-world simulation, must also be driven by chain-of-thought reasoning. Concordia addresses this need by providing code abstractions that facilitate the integration and management of these dynamic environmental states and their reasoned evolution.

2.5 Challenges of Using LLM-Derived Synthetic Personas

As discussed by Kapania et al. (2024), there are a number of challenges in using LLMs as synthetic personas (users). LLMs are a consolidation over human experience, and hence they combine multiple viewpoints, failing to represent the diversity of human experience or correlations between different aspects of individuals. They can also lack depth: for instance, they can make references to particular phenomena (e.g., sleep difficulties in our test domain), but these phenomena do not represent situated knowledge, grounded in lived experience and history. This makes it difficult to model human users, since the relevant reasons for the health conditions might be missing. We mitigate this issue with the option of adding backstories, which have been shown to provide grounding to LLM-generated output (Moon et al., 2024).

Training data may also be biased as they were drawn largely from English-speaking cultures and individuals with strong online presence. To demonstrate the issue in the domain of sleep disorders, a survey of 27,000 US adults revealed that difficulty in falling and in staying asleep occurs more commonly among residents in non-metropolitan areas, and for those on lower incomes (Adjaye-Gbewonyo et al., 2022). We address this bias by specifying persona statistics drawn from relevant demographic health studies (e.g., Yfantidou et al., 2022; Arges et al., 2020), rather than relying on generation by the LLM.

A further subtle risk of using LLMs in simulating synthetic users can be understood from a causal inference perspective, as discussed in Gui and Toubia (2023). Providing particular advice to an LLM (e.g., a recommendation regarding better sleep practices) can inadvertently cause unintended variation in other factors (e.g., the LLM's beliefs about the age and lifestyle factors of the synthetic persona). We mitigate this effect through explicit control over demographic and personality factors that the LLM might otherwise incorrectly impute, a practice recommended by Gui and Toubia (2023), although this remains an ongoing research challenge.

3 Methods

While LLMs trained on a large data corpus (e.g., the entire Internet) partially represent the general human population, the distributions of subpopulations and health conditions represented in a single model are skewed and poorly understood (Kapania et al., 2024). Furthermore, the instruction fine-tuning step (e.g., reinforcement learning with human feedback) typically employed by high-performance LLMs further impacts the distribution of generated outputs. In this work we propose an end-to-end framework for generating synthetic users grounded in reality, by starting from human user data including demographics, health & lifestyle, and behavioral & psychological characteristics, to accurately represent a desired population (Figure 1). This real cohort can be sampled either uniformly, or with oversampling or undersampling explicitly designed by researchers (e.g., to better surface underrepresented subgroups or conditions). Given this sampled cohort, we optionally add additional health conditions conditioned on already specified attributes of the synthetic users, either by using the LLM or via a rule-based algorithm, and also optionally add rich backstories of the synthetic user, conditioned on existing attributes (Moon et al., 2024).

This combination of background information grounded in real data constitutes a "vignette" of the synthetic user, all based in natural language. In the following sections we explore how these vignettes can be used to generate interactions between synthetic users and coaching agents, in two separate health coaching scenarios–sleep coaching and diabetes coaching–with two independently developed coaching agents.

4 Experiments: Sleep Coaching

Sleep is vital for health and well-being; insufficient sleep and untreated sleep disorders can have a detrimental effect on cognitive function, mood, mental health, and cardiovascular, cerebrovascular, and metabolic health (Ramar et al., 2021). Sleep disorders are highly prevalent: according to the 2020 National Health Interview Survey (NHIS) (Adjaye-Gbewonyo et al., 2022), 14.5% of US adults had trouble falling asleep most days over the 30-day study period, and over a quarter of adults do not meet the minimum recommended sleep duration per night.

4.1 Synthetic Users with Sleep Conditions

To generate synthetic users grounded in real sleep conditions, we utilized the publicly available LifeSnaps dataset, a multi-modal, longitudinal, geographically-distributed dataset containing participant demographics, smartwatch measurements including sleep data, health, behavioral, and psychological trait surveys, and ecological momentary assessments (Yfantidou et al., 2022). Among the fields available in this dataset, we focused on basic demographics (age, gender), basic health & sleep attributes (body mass index, sleep duration & efficiency), and five personality markers from the International Personality Item Pool (IPIP) version of the Big Five (Goldberg, 1992). Importantly, given longitudinal sleep data over four months in LifeSnaps, we included both the average sleep duration and the variability of sleep duration, both important health factors. See Appendix A for more details.

As outlined in Figure 1, we generated a synthetic "sleep profile" for each individuals in LifeSnaps (N=68) conditioned on real demographics, sleep duration & efficiency, and behavioral characteristics. The generated sleep profile consists of the following four key attributes: "primary sleep concern", "sleep goals", "reasons for goals", and "barriers". The full prompt used to generate these attributes is available in Appendix C. This sleep profile and all aforementioned structured attributes from LifeSnaps formed the "vigenettes" for our synthetic users with sleep conditions.

Finally, we instantiated each synthetic user in Concordia (Vezhnevets et al., 2023) as a "SimpleLLMAgent" to prioritize clarity and reproducibility, where the entire conversation history and the synthetic user's backstory are included within the prompt context for generating each subsequent utterance (Appendix B) [1].

4.2 The Sleep Coaching Agent

We used a sleep coaching agent based on Christakopoulou et al. (2024), which proposed a two-agent system consisting of a "Talker" and a "Reasoner" agent. Inspired by Kahneman (2011), the

[1]A code snippet for instantiating our synthetic users using Concordia is available at `https://anonymous.4open.science/r/sleepless_nights`

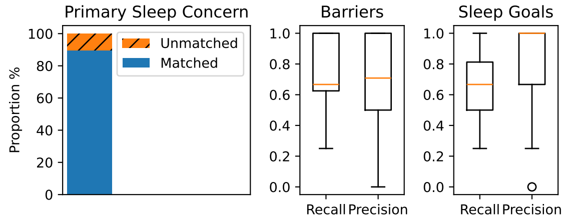

Figure 2: **Extracting synthetic users' sleep profiles from interaction for automated evaluation of the coaching agent**. Our synthetic users (N=68) with sleep conditions interacted with our sleep agent for 10 turns. Recall and precision were computed by comparing the internal state-of-user model in the coaching agent to the true sleep profile of the synthetic user. "Primary sleep concern" is a single natural language statement, while "barriers" and "sleep goals" include multiple natural language items.

Reasoner ("System 2") is responsible for generating an internal model of the use based on conversation history, in addition to planning and calling tools, while the Talker ("System 1") is responsible for generating conversation based on the structured state generated by the Reasoner. We used Gemini 1.5 Pro (Gemini Team, 2024) as the LLM engine for both constituent agents.

4.3 Evaluating Agent-Synthetic User Interactions in Sleep Coaching

Explicit modeling of the user in the sleep coaching agent enabled us to compare the internal user state to the "true" sleep profile we explicitly assigned to the synthetic user. After a 10-turn interaction between a synthetic user and the coaching agent (the turn length of 10 was chosen to balance efficiency and comprehensiveness, after reviewing real sleep coach-user interactions in Perlis et al. (2005)), the coaching agent was able to identify the synthetic user's primary sleep concern with 89.7% accuracy (Figure 2; error analysis in Appendix P). For barriers and sleep goals, which are also explicitly modelled by the coach agent and consist of multiple items (natural language sentences) per user, we computed recall (sensitivity) and precision, as measured by the proportion of the true attributes also captured by the coaching agent's model, and the proportion of the attributes in the coaching agent's model that were present in the true attributes, respectively. When matching sentences, we used fuzzy-matching by a high-performance, instruction-tuned LLM (Gemini 1.5 Pro; Gemini Team (2024)), to account for paraphrasing (e.g., "inconsistent sleep duration" and "variable sleep duration"; Appendix D), and manually checked a random subset (10 individuals) to make sure the answers were consistent with human answers (an easy task for modern LLMs; see, for example, Michail et al. (2025)). We obtained 71.4% mean recall and 72.5% mean precision for the barriers, and 66.4% mean recall and 84.2% mean precision for the sleep goals (Figure 2). We also conducted the same experiment using Gemma 2 ("27B-IT" version) (Gemma Team, 2024) as an alternative open-source backend LLM for synthetic user generation in Concordia, and observed performance degradation (Appendix O).

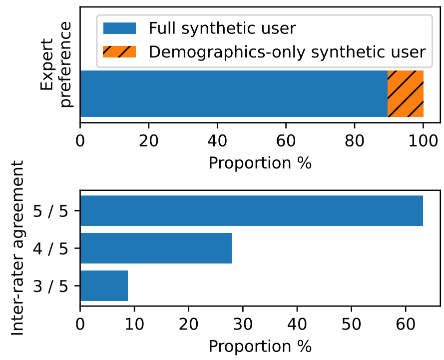

Figure 3: **Expert evaluation of grounded synthetic users**. Human expert evaluation of interactions from two sets of synthetic users, one from the full synthetic user pipeline and another from demographics-only data without sleep conditions or behavioral/psychological characteristics, with a fixed sleep coaching agent. Five human experts annotated each case independently.

Finally, to validate our use of synthetic users with health and behavioral/psychological attributes in addition to basic demographic information, we conducted a human expert evaluation of interactions from two synthetic users: one from our full synthetic user pipeline and another from using demographics-only data without health conditions, with the same sleep coaching agent. We asked human evaluators, with training and quality control conducted by a clinical psychologist, to annotate their preference between two interactions with each synthetic user, with 5$\times$ coverage of each question (Appendix R). The evaluators overwhelmingly favored our full synthetic users over the baseline and the differences were highly statistically significant (p-value $= 3.7 \times 10^{-12}$ by one-tailed test under binomial null distribution), with high inter-rater reliability (64% of the cases achieving the complete 5/5 agreement and 91% of the cases achieving 4/5 agreement or more; Fleiss' $\kappa = 0.67$ (Fleiss, 1971)) (Figure 3).

5 Experiments: Diabetes Coaching

Diabetes mellitus, commonly referred to as diabetes, is a highly prevalent chronic disease, affecting 15% of all US adults in 2021 (CDC, 2024). Understanding lifestyle barriers is crucial for effectively addressing the challenges this condition presents (Deslippe et al., 2023). The barriers significantly impact an individual's ability to adhere to recommended behaviors and achieve optimal health outcomes. Many barriers are intricately driven by demographic, socioeconomic, and clinical circumstances, collectively manifesting into specific challenges. By accurately representing these interconnections, we can create synthetic user profiles that faithfully reflect the real-world complexities encountered by individuals with cardiometabolic conditions, thereby enhancing the realism and applicability of our synthetic users interacting with a diabetes coaching agent (Figure 1).

5.1 Synthetic Users with Diabetes

To obtain a representative distribution of cardiometabolic barriers, we leveraged insights from 100 peer-reviewed articles in Yang et al. (2024), which comprehensively identified 246 challenges encountered by cardiometabolic patients, in consultation with behavioral experts. The challenges were classified into six distinct sub-categories within the COM-B (capability, opportunity, motivation - behavior) model, a widely recognized framework for understanding behavior in healthcare (Michie et al., 2011). The challenges were again categorized into 21 distinct barrier concepts by behavioral experts, each belonging to a sub-category within the COM-B model, as shown in Appendix E. Through this process, we ensured that the synthetic user profiles developed for simulated coaching agent interactions were representative, reflecting the appropriate distribution of challenges.

To derive our synthetic users with diabetes, we utilized Project Baseline Health Study (PBHS), a longitudinal cohort with diverse backgrounds representative of the entire health spectrum (Arges et al., 2020).[2] After careful preprocessing (Appendix G), we obtained a cohort of 345 Type 2 diabetic individuals with diverse demographic, social, medical, and health attributes (Appendix H).

To construct realistic vignettes of synthetic users from this cohort, we first sampled a COM-B category under its original distribution in PBHS (Appendix F), and then uniformly sampled a specific barrier within the selected COM-B category. Using this identified barrier, we randomly selected a corresponding individual from PBHS with relevant symptoms. The selected individual's demographic, socioeconomic, clinical, and behavioral survey data formed the foundation of our natural language vignette (Figure 1). To enhance the narrative quality, realism, and depth of these vignettes, a coherent backstory was generated by an LLM, incorporating the identified barrier (Appendix J). Finally, we generated a communication style field (e.g., tone, verbosity, and confidence) to compose the final vignette. Notably, the specific technical term for the selected barrier was not included in the vignette, since human users are unlikely to label themselves under it. In total, 200 vignettes for synthetic users were generated using the commercially available Gemini 2.0 Flash model chosen for its reasoning and conversation capabilities (Pichai, 2024).

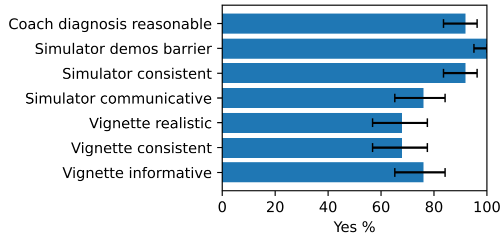

Figure 4: **Assessing synthetic users' performance from interaction with the diabetes coaching agent**. Three human annotators reviewed 25 randomly-selected simulated interactions. Wilson's confidence interval was used for these binary annotation tasks.

5.2 The Diabetes Coaching Agent

The synthetic users generated from the vignettes interacted with a diabetes coaching agent, building upon methodologies developed by Yang et al. (2024). This coaching agent was constructed to assist users to set health goals and overcome identified barriers. After each simulated interaction, the coaching agent was asked to identify one of the 21 barriers in its modeling of the user. Both the synthetic users and the coaching agent were generated using Gemini 1.5 Pro (Gemini Team, 2024).

5.3 Evaluating Agent-Synthetic User Interactions in Diabetes Coaching

To assess the fidelity and reliability of the synthetic users, we solicited evaluations from a panel of three experts, comprising behavioral scientists

[2] The data use was formally approved by our institution's internal review committee, ensuring ethical compliance.

and patient care practitioners (Appendix K). Each expert received 25 randomly-selected simulated interactions (out of 200) for assessment. Experts agreed that the synthetic users were highly consistent (92%) and effectively demonstrated the barriers they were designed to portray (100%) (Figure 4; Appendix L).

Similarly to our sleep coaching evaluation experiments (Section 4.3), we also conducted a secondary experiment comparing our synthetic users to baseline synthetic users generated from general demographic data only, without the PBHS data and the COM-B barriers that our methodology uniquely provides, in their interactions with the same coaching agent. After randomly selecting an individual from PBHS, the baseline synthetic user was generated from their standard demographic information, while our full synthetic user had the same standard demographics, richer health data, and the individual's original barrier (Appendix I). Both synthetic users were asked to portray an individual with cardiometabolic health challenges, and we ensured both possessed the same standard demographic data, communication style, verbosity, and confidence, for a fair comparison.

Human annotators were assigned three evaluation questions (Appendix M) to annotate randomly selected interactions without actual vignettes, testing implicit demonstration of the vignettes through conversations. 75 pairs of interactions (full synthetic users and baseline synthetic users) were evaluated (Figure 5).

For consistent representation of a single barrier (correct or not), annotators preferred the interactions from our full synthetic users twice as much as the baseline (32%:15%). Moreover, the annotators overwhelmingly preferred the full synthetic user (70%) for demonstrating the *correct* original barrier. We observed mixed results for informativeness, with the baseline being slightly preferred and more than half of the interaction pairs annotated as "Similar". Overall, this analysis demonstrated that our synthetic users can more accurately portray the original barrier consistently in simulated interactions (Figure 5), which has an important downstream effects. This implies our synthetic user pipeline can be used to generate training data for efficiently fine-tuning the coaching agent for a specific barrier for incremental improvement. On the other hand, simulated interactions with the baseline synthetic users would not reliably represent the specific barrier one seeks to better capture.

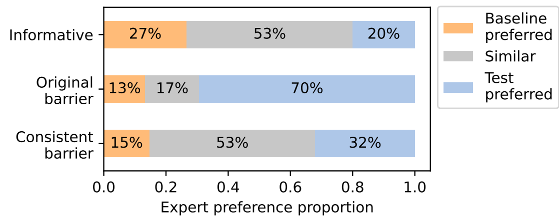

Figure 5: **Comparing grounded synthetic users to a baseline on simulated diabetes coaching conversations.** Human annotators reviewed 75 pairs of conversations, and labeled which conversation demonstrated the dimensions represented in each row. "Test" represents our grounded synthetic users and "baseline" represents synthetic users created from basic demographics only.

Finally, we investigated the difference in distribution of barriers that the baseline synthetic users (using only demographic information) created, compared to the reference distribution observed in the literature (Appendix N). As the baseline synthetic users only express barriers through conversation, we used the coaching agent's internal state ("diagnosis") of the user's barrier established through interaction, which was evaluated as credible by human experts (Figure 4). Interestingly, the baseline synthetic users significantly over-sampled certain barriers (physical capability, reflective motivation), while under-sampling others (social opportunity, physical opportunity), highlighting the importance of grounding synthetic users in behavioral data from real human users whom the coaching agent is designed to serve.

6 Conclusion

We designed an end-to-end framework for generating synthetic users for evaluating coaching agents grounded in health, lifestyle, behavioral, and psychological attributes complementing basic demographics from human data. The additional sampling process was employed to ensure that the distribution of these attributes could be explicitly controlled (i.e., to ensure applicability to less represented sub-populations). Realistic backstories and additional health conditions (where not available in human user data) were conditionally generated from the grounded data. The final vignettes generated by this framework were used in a generative agent-based model framework to effectively simulate interactions between the synthetic users and the given coaching agent. We evaluated our methods in two independent, highly prevalent health coaching use cases of sleep coaching and diabetes management.

Efficient development of autonomous agents is significantly accelerated by evaluation metrics that can be computed *without* complete dependence on human users, since human evaluation is costly and time-consuming. While there is a large body of literature on the design of general-purpose synthetic users in this setting (see Section 2), we have highlighted the importance of grounding synthetic users' health and behavioral attributes on a real human dataset, with concrete demonstrations and evaluations of our end-to-end pipeline in two independently developed coaching agents. Although some of our demonstrations relied on access to the coaching agents' internal user model, the simulated interactions generated by our framework can be evaluated in a way that is agnostic to the design of the agent, so the framework can be applied to any conversational agents.

The extensive, realistic, and grounded simulated interactions produced by our framework lay the foundation for efficient development of coaching agents, potentially through a fine-tuning or reinforcement learning loop, as will be pursued in future research.

Limitations

This work introduced an end-to-end framework for evaluating a coaching agent using automated synthetic users. The ultimate goal of performing this evaluation is to improve the coaching agent itself using the metrics and signals derived from this work, which is left for future research. The coaching agents we evaluated in this work may not have the "state-of-the-art" performance, as that was not the focus of this paper. We also did not compare multiple commercial LLMs, as that was also not the focus of this paper.

Using LLM-powered agents presents risks for users. One of the motivations for this work is characterizing and minimizing some of those risks by enabling predictive agent evaluation with realistic user personas.

In evaluating coaching agents, a longitudinal assessment of potential behavioral change is crucial. Effectively and realistically simulating synthetic users' behavioral change is a critical open question to be addressed, including modeling non-adherence to coaching advice (e.g., human users may not follow the advice, or may *claim* they changed their behavior, even if they did not).

Personalization and adaptation of coaching and other agents are very important and relevant topics that are beyond the scope of this paper. We hope that grounded, rich, and diverse personas of synthetic users we developed in this work will significantly support the process of evaluating and improving personalized agents adapting to evolving user behaviors and needs.

References

Dzifa Adjaye-Gbewonyo, Amanda E. Ng, and Lindsey I. Black. 2022. Sleep difficulties in adults: United states, 2020. *NCHS Data Brief*, (436).

Kristine Arges, Themistocles Assimes, Vikram Bajaj, Suresh Balu, Mustafa R Bashir, Laura Beskow, Rosalia Blanco, Robert Califf, Paul Campbell, Larry Carin, Victoria Christian, Scott Cousins, Millie Das, Marie Dockery, Pamela S Douglas, Ashley Dunham, Julie Eckstrand, Dominik Fleischmann, Emily Ford, Elizabeth Fraulo, John French, Sanjiv S Gambhir, Geoffrey S Ginsburg, Robert C Green, Francois Haddad, Adrian Hernandez, John Hernandez, Erich S Huang, Glenn Jaffe, Daniel King, Lynne H Koweek, Curtis Langlotz, Yaping J Liao, Kenneth W Mahaffey, Kelly Marcom, William J Marks, Jr, David Maron, Reid McCabe, Shannon McCall, Rebecca McCue, Jessica Mega, David Miller, Lawrence H Muhlbaier, Rajan Munshi, L Kristin Newby, Ezra Pak-Harvey, Bray Patrick-Lake, Michael Pencina, Eric D Peterson, Fatima Rodriguez, Scarlet Shore, Svati Shah, Steven Shipes, George Sledge, Susie Spielman, Ryan Spitler, Terry Schaack, Geeta Swamy, Martin J Willemink, and Charlene A Wong. 2020. The project baseline health study: a step towards a broader mission to map human health. *NPJ Digit. Med.*, 3(1):84.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. 2024. Persona: A reproducible testbed for pluralistic alignment. *Preprint*, arXiv:2407.17387.

CDC. 2024. National diabetes statistics report. `https://www.cdc.gov/diabetes/php/data-research/index.html`. Accessed: 2025-2-11.

Konstantina Christakopoulou, Shibl Mourad, and Maja Matarić. 2024. Agents thinking fast and slow: A talker-reasoner architecture. *Preprint*, arXiv:2410.08328.

Alysha L Deslippe, Alexandra Soanes, Celeste C Bouchaud, Hailee Beckenstein, May Slim, Hugues Plourde, and Tamara R Cohen. 2023. Barriers and facilitators to diet, physical activity and lifestyle behavior intervention adherence: a qualitative systematic review of the literature. *Int. J. Behav. Nutr. Phys. Act.*, 20(1):14.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychol. Bull.*, 76(5):378–382.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *arXiv [cs.CL]*.

Lewis R Goldberg. 1992. The development of markers for the big-five factor structure. *Psychological assessment*, 4(1):26.

Google. 2023. Palm 2 technical report. *Preprint*, arXiv:2305.10403.

George Gui and Olivier Toubia. 2023. The challenge of using llms to simulate human behavior: A causal inference perspective. *SSRN Electronic Journal*.

Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Sci. Data*, 3(1):160035.

Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Leandra A Barnes, Hong-Yu Zhou, Zhuo Ran Cai, Eliezer M Van Allen, David Kim, Roxana Daneshjou, and Pranav Rajpurkar. 2025. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nat. Med.*, 31(1):77–86.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Shivani Kapania, William Agnew, Motahhare Eslami, Hoda Heidari, and Sarah Fox. 2024. 'simulacrum of stories': Examining large language models as qualitative research participants. *Preprint*, arXiv:2409.19430.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024. RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. *Preprint*, arXiv:2309.00267.

Megan M MacPherson and Shabana Kapadia. 2023. Barriers and facilitators to patient-to-provider messaging using the COM-B model and theoretical domains framework: a rapid umbrella review. *BMC Digit Health*, 1(1).

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *Preprint*, arXiv:2303.17651.

Andrianos Michail, Simon Clematide, and Juri Opitz. 2025. PARAPHRASUS: A comprehensive benchmark for evaluating paraphrase detection models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8749–8762.

Susan Michie, Maartje M van Stralen, and Robert West. 2011. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. *Implement. Sci.*, 6(1):42.

Suhong Moon, Marwa Abdulhai, Minwoo Kang, Joseph Suh, Widyadewi Soedarmadji, Eran Kohen Behar, and David Chan. 2024. Virtual personas for language models via an anthology of backstories. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. Generative agent simulations of 1,000 people. *Preprint*, arXiv:2411.10109.

Michael L Perlis, Carla Jungquist, Michael T Smith, and Donn Posner. 2005. *Cognitive behavioral treatment of insomnia: A session-by-session guide*. Springer, New York, NY.

Sundar Pichai. 2024. Introducing gemini 2.0: our new AI model for the agentic era. `https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/`. Accessed: 2025-2-13.

Kannan Ramar, Raman K Malhotra, Kelly A Carden, Jennifer L Martin, Fariha Abbasi-Feinberg, R Nisha Aurora, Vishesh K Kapur, Eric J Olson, Carol L Rosen, James A Rowley, Anita V Shelgikar, and Lynn Marie Trotti. 2021. Sleep is essential to health: an american academy of sleep medicine position statement. *J. Clin. Sleep Med.*, 17(10):2115–2119.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *Preprint*, arXiv:2307.00184.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Yong Cheng, Le Hou, Albert Webson, Kavita Kulkarni, S Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan. 2024. Towards conversational diagnostic AI. *Preprint*, arXiv:2401.05654.

Alexander Sasha Vezhnevets, John P Agapiou, Avia Aharon, Ron Ziv, Jayd Matyas, Edgar A Duéñez-Guzmán, William A Cunningham, Simon Osindero, Danny Karmon, and Joel Z Leibo. 2023. Generative agent-based modeling with actions grounded in physical, social, or digital space using concordia. *arXiv [cs.AI]*.

Xiaoyang Wang, Hongming Zhang, Tao Ge, Wenhao Yu, Dian Yu, and Dong Yu. 2025. Opencharacter: Training customizable role-playing llms with large-scale synthetic personas. *Preprint*, arXiv:2501.15427.

Yan Wang, Heidi Donovan, Sabit Hassan, and Malihe Alikhani. 2023. MedNgage: A dataset for understanding engagement in patient-nurse conversations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4613–4630, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Eric Yang, Tomas Garcia, Hannah Williams, Bhawesh Kumar, Martin Ramé, Eileen Rivera, Yiran Ma, Jonathan Amar, Caricia Catalani, and Yugang Jia. 2024. From barriers to tactics: A behavioral science-informed agentic workflow for personalized nutrition coaching. *Preprint*, arXiv:2410.14041.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.

Sofia Yfantidou, Christina Karagianni, Stefanos Efstathiou, Athena Vakali, Joao Palotti, Dimitrios Panteleimon Giakatos, Thomas Marchioro, Andrei Kazlouski, Elena Ferrari, and Sarunas Girdzijauskas. 2022. LifeSnaps, a 4-month multi-modal dataset capturing unobtrusive snapshots of our lives in the wild. *Sci. Data*, 9(1):663.

Huizi Yu, Jiayan Zhou, Lingyao Li, Shan Chen, Jack Gallifant, Anye Shi, Xiang Li, Wenyue Hua, Mingyu Jin, Guang Chen, Yang Zhou, Zhao Li, Trisha Gupte, Ming-Li Chen, Zahra Azizi, Yongfeng Zhang, Themistocles L. Assimes, Xin Ma, Danielle S. Bitterman, Lin Lu, and Lizhou Fan. 2024. Aipatient:

Simulating patients with ehrs and llm powered agentic workflow. *Preprint*, arXiv:2409.18924.

Yue Zhou, Barbara Di Eugenio, Brian Ziebart, Lisa Sharp, Bing Liu, and Nikolaos Agadakos. 2024. Modeling low-resource health coaching dialogues via neuro-symbolic goal summarization and text-units-text generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11498–11509, Torino, Italia. ELRA and ICCL.

A LifeSnaps Dataset Preprocessing

The age and the gender of the participants in the LifeSnaps dataset were available in two categories: “under 30” and “over 30”, and “male” and “female”. Of the 71 participants, we excluded 3 individuals whose age or gender data were missing in the dataset, resulting in 68 individuals. BMI was classified into four categories following a standard practice and terms: “underweight” (BMI < 19), “normal” ($19 \leq$ BMI < 25), “overweight” ($25 \leq$ BMI < 30), and “obese” (BMI ≥ 30).

Given longitudinal LifeSnaps data over four months, we computed the mean and the standard deviation of sleep duration. The mean sleep duration was included in the vignettes in the number of hours. Participants with whose standard deviation of duration was no more than 1.5 hours were considered to have “consistent” sleep duration, and the rest were considered to have “variable” sleep duration.

B Synthetic User Prompt Generation for the Sleep Domain

Prompt components: `Vignette`, `Context`, and `Instructions`. A `Context` consists of the synthetic user’s `Name` and the entire `conversational transcript` so far.

```
Nicole is female. She is more than 30 years old. She typically sleeps for 8
hours at night. Her sleep duration is highly variable. Her sleep efficiency
is typically at 94%. She is underweight. She is neither introverted nor
extraverted. She is highly agreeable. She is moderately conscientious. She
feels unstable. She is highly intellectual. (...)
You are Nicole.
Your last observations are:
Nicole - “I’m struggling with getting enough sleep. I’d like to sleep 7-8
hours a night consistently so I can improve my energy levels at work and
be more present for my family, but my demanding work schedule and racing
thoughts make it difficult.”
COACH - “It’s completely understandable that a demanding work schedule and
racing thoughts can make sleep a challenge. Wanting better energy and to be
more present for your family are fantastic goals! Is there something specific
you’ve tried in the past to help with sleep, or something you’d like to try
now?” (...)
Given the above, generate what Nicole would say to the COACH next in this
conversation. Respond in the format “Nicole - ...”. If the COACH has asked
for input, make sure to generate specific and detailed answer to the question.
Bear in mind the primary sleep concern, sleep goals, reasons for those goals,
and barriers of Nicole provided above. The tone and style of the conversation
should math Nicole’s descriptions above.
```

C Sleep Profile Generation

The following prompt was used to generate sleep profiles for synthetic users based on LifeSnaps, where [name] is randomly sampled and [background_info] is generated by real data.

> Here is some background information about a fictional person, called [name]. I want to create a fictional profile of [name]. The sleep profile should be in a structured JSON format, including these fields: "primary_sleep_concern", "sleep_goals", "reasons_for_goals", "barriers". "primary_sleep_concern" must be a single string. "sleep_goals" must be a list of strings. "reasons_for_goals" must be a list of strings. "barriers" must be a list of strings. Be creative and make up stories if you need to, but make sure the JSON sleep profile is consistent with the given background information. Your response must start with a justification of each field you create, followed by the structured JSON. It’s important that you only use escaped double quotes and single quotes, as well as escaped backslashes to not break the JSON format. It’s also important that you respect the JSON format.

Background information: "[background_info]"

Response:

D Prompts for Fuzzy Matching Sleep Profiles

For fuzzy-matching two descriptions of primary sleep concerns to account for paraphrasing, the following prompt was used for an instruction-tuned high-performance LLM.

> I would like you to answer this question.
>
> The following are two descriptions of someone's primary sleep concern, A and B:
>
> A: "[description A]"
>
> B: "[description B]"
>
> Does A and B generally communicate compatible sleep concerns? They do not have to match completely, but I want to know whether they GENERALLY describes similar concerns. It is okay if some specific details are different.
>
> Start your answer with Yes or No, followed by an explanation.

For fuzzy-matching a list of sleep attributes (such as sleep goals and barriers), the following prompt was used to generate precision and recall.

> I would like you to answer two questions. I have two lists of [field name] of a person. I want to know how they match well with each other. There are [length of list A] items in the list A, and there are [length of list B] items in the list B:
>
> A: [list A]
>
> B: [list B]
>
> My two questions are: How many of the items in the list A are also generally represented in the list B? And how many items in the list B are also generally represented in the list A?
>
> When finding a match, not all the details need to match. I just want to know whether each item is generally represented in the other list, while some specific details may be different.
>
> Start your answer with two numbers separated by a comma and followed by a period, representing the answers to the two questions, for example "3, 2.". Add your explanation after this answer.

E COM-B Categories

COM-B Category	Barriers
Psychological Capability	Don't know the Basics, Don't know the consequences, Planning fallacy, Memory
Physical Capability	Decision fatigue, Physical limitations
Social Opportunity	Lack of social support, Conflicting opinions, Impact on others, peer pressure
Physical Opportunity	Geographic limitations, Affordability or costs, Lack of equipment, Switching settings
Reflective Motivation	Poor self-efficacy, Competing priorities , Lack of desire without reasons, Boredom
Automatic Motivation	Present bias, Anchoring effect, Gut feelings

F COM-B Category Distribution

The COM-B sub-category distribution was 25% reflective motivation, 21% psychological capability, 19% physical opportunity, 15% social opportunity, 12% automatic motivation, 9% physical capability. This relative distribution of the barriers were also observed in related studies (MacPherson and Kapadia, 2023).

G PBHS Diabetes Data Preprocessing

First, we filtered the PBHS dataset to include only individuals diagnosed with type 2 diabetes, resulting in a cohort of 345 individuals. For each individual, we leveraged their comprehensive demographic, socioeconomic, clinical, and behavioral survey data. The data encompassed a range of variables, including age, sex, race, marital status, smoking status, education level, income, insurance coverage, household size and structure, social interactions (measured through gatherings, phone calls, and text messages), attendance at organizational meetings, and employment status. Additionally, clinical parameters such as diagnosed conditions, Hemoglobin A1c (HbA1c) levels, blood glucose measurements, (systolic/diastolic) blood pressure, and body mass index (BMI) were used. Furthermore, the survey data underwent manual coding to identify if any of the individuals reported experiencing any relevant symptoms of the 21 identified barriers as their primary challenges. This coding process ensured that the nuanced, self-reported experiences of the individuals were captured and integrated into the development of our simulated user profiles.

H PBHS Diabetes Data Attributes

The following fields from PBHS data were used as input for synthetic user vignette generation (see Appendix J.

Data Category	Attribute
Basic demographics	Age at enrollment
Basic demographics	Race
Basic demographics	Marital status
Basic demographics	Education
Basic demographics	Income
Basic demographics	Employment status
Social environment	People living at home
Social environment	Number of people under 18 living at home
Social environment	Weekly number of friend and family gatherings
Social environment	Weekly number of phone calls with friends and family
Social environment	Weekly number of texts with friends and family
Social environment	Weekly attendance to social organization meetings
Health data	Smoking status
Health data	Has insurance
Health data	Diagnostic conditions
Medical measurements	Hemoglobin A1C (HbA1c)
Medical measurements	Blood Glucose (mg/dl)
Medical measurements	Systolic and diastolic blood pressure
Medical measurements	Body Mass Index (BMI)
Barrier	Matched barrier from PBHS data

I Synthetic User Prompt for the Diabetes Domain

Prompt components: `Data`, `Vignette`, and `Instructions`. `Data` is the synthetic user's demographic, social, and medical data from PBHS. `Vignette` is the backstory of synthetic user's barriers and `Data` summarized in a few sentences. Conversation history is tracked and accounted for separately through a stateful chat session to ensure synthetic user is consistent.

```
You are a patient with cardiometabolic condition using a digital chat based
application, below is your specific role. Pursue the conversation and the
AI Coach will work with you. Reference your patient details and bring up
information that can be relevant to the conversation. ONLY write out the
conversation (not the breathing nor internal thoughts).
```

```
Patient details: [PBHS data (..)]

John is a 31-year-old working dad, with a busy life. He lives with his partner
and three young children and is trying to juggle work and family life. He
knows he needs to eat better to manage his diabetes with an HbA1c of 6.7 and a
high blood glucose level of 305, and he genuinely wants to improve. He often
tells himself he will eat a healthy lunch, but finds himself eating fast food
because he didn't have time to pack a lunch or get the right groceries the
night before. He knows he's not doing his best, and feels like he is letting
himself down, but also feels overwhelmed and unsure of where to even begin.

Now it's your turn to speak, you are the patient. You should follow your role
and continue the conversation based on the AI Coach's last message.
- Make sure you focus on the backstory section of the patient details -
specifically the BARRIER the patient is facing.
- Use the other patient details as supporting information in a natural, smooth
way if necessary. When using those details, make sure to talk like a regular
patient (change medical jargon to common speak) and not a doctor.
- DO NOT hallucinate or make up any new information that are not in the
patient details. Stick to the context in the backstory presenting the barrier
/ challenge.
- Make sure to follow the communication style given in the patient details.
- You should only respond with at most 2 sentences per turn.
```

J Diabetes Synthetic User Vignette Generation

J.1 Initial Vignette Generation

Prompt components: `Data`, `Barrier`, `Few Shot Examples`, and `Instructions`. `Data` is the synthetic user's demographic, social, and medical data from PBHS. `Few Shot Examples` are example synthetic user vignettes in JSON format to guide vignette generation. `Barrier` is the barrier the synthetic user is facing randomly selected from a prior distribution. The prompt below is for our test vignette; for the baseline vignette, we exclude health data and attributes unique to PBHS.

```
You are an expert in creating realistic patient vignettes for a digital
health app with an AI coach providing goals achievement support for patients
with diabetes. You are tasked with creating a realistic vignette for a
patient working on their health goals. Given a barrier that a patient is
experiencing, fill in the the following possible parameters that are aligned
to each other and to the barrier, realistic for a diabetes patient, and
collectively and informatively represent the barrier.

Here is the barrier the patient is experiencing: [Barrier]

Here are the possible parameters you can fill in: [PBHS Data (...)]

Include the following additional parameters as well:
- Name: Patient's name matching the patient SEX provided.
- Tone: (formal, academic, casual, playful, agreeable, antagnoistic,
resistant, depressed, apathetic)
- Verbosity: (Shares intentional, complete sentences, responds to each
question asked; Responds in short sentences or phrases; Shares unrelated
```

```
information / overshares)
- Confidence: (High confidence and self awareness, knows themselves and conveys accurately; Concerned for appearance, erring towards aspirational self, overly optimistic view of oneself; Low confidence, convinced they are likely doing something wrong, apologetic)

Here are the rules for creating the vignette:
- First, fill in each field of the parameters with a realistic value that is aligned with the barrier. The values can be numerical values or short free text descriptions.
- Then, provide a short backstory paragraph that describes the patient's situation, including the filled in parameters. The paragraph should be realistic and informative to represent the barrier.
- IMPORTANT: It is critical that the specific barrier term is not used in the vignette, including the backstory. The vignette should represent the barrier without explicitly mentioning it.

 Here is an example of a vignette with the barrier of [Barrier]:
[Few shot examples]


Your final output should be the following, in JSON format.
{{
"reasoning": <Your reasoning behind the vignette you are generating>,
"vignette": <The vignette you have created in valid python dictionary format with keys containing the relevant parameter fields with an additional key containing the overall backstory.>
}}
Think step by step, and validate your reasoning with your text.
```

J.2 Vignette Generation Improvement

After initial generation, we developed another "verifier LLM" to improve the quality of the vignettes. This verifier LLM edited components of the vignette and backstory to ensure that they were realistic and consistent with each other. This idea of "self-refinement" was inspired by existing literature (Madaan et al., 2023; Yao et al., 2023), but we used a separate LLM with fewer iterative steps. The prompt used by our verifier LLM is shown below.

Prompt components: `Vignette`, `Barrier`, `Few Shot Examples`, and `Instructions`. `Vignette` is the initially generated synthetic user vignette to improve. `Few Shot Examples` are example synthetic user vignettes in JSON format to guide vignette generation. `Barrier` is the barrier the synthetic user is facing randomly selected from a prior distribution.

```
You are an expert in realistic patient vignettes for a digital health app with an AI coach providing goals achievement support for patients with diabetes. You are tasked with improving / editing an existing realistic patient vignette provided to you.

Use the following methods to improve the vignette and parameters:
1) Remove information that causes the vignette to be unrealistic or misaligned with the barrier.
2) Ensure the vignette is representative to ONLY the barrier provided to you, without explicitly mentioning the barrier.
3) Add information that makes the vignette more realistic.
```

```
4) Double check that the backstory is using the correct age field and not
hallucinating.
5) Remove any unnecessary formatting (curly brackets, hallucinations, etc).

Here is the barrier the patient is experiencing: [Barrier]

Your final output should be the following, in JSON format.
{{
"reasoning": <Your reasoning behind how you improved the vignette>,
"vignette": <The vignette you have improved in valid python dictionary
format with the exact same keys as the provided vignette.>
}}
Think step by step, and validate your reasoning with your text.

Improve the following vignette: [Vignette]

EXAMPLES:
[Few Shot Examples (...)]
```

K Diabetes Evaluation

Human experts were asked to provide insights by responding to the following evaluative questions:

1. The vignette informatively represents the specified barrier. (Y/N)
2. The information in the vignette is self-consistent. (Y/N)
3. The traits in the vignette are realistic for a cardiometabolic patient. (Y/N)
4. The patient simulator's communication style specified in the vignette is reflected in the conversation. (Y/N)
5. The patient simulator's responses are consistent with the vignette. (Y/N)
6. The patient simulator informatively demonstrates the barrier specified in the vignette. (Y/N)
7. The coaching agent's diagnosis of patient barriers is reasonable given the conversation alone. (Y/N)

L Diabetes Evaluation Results

Question	Expert 1 Yes Response Rate	Expert 2 Yes Response Rate	Expert 3 Yes Response Rate	All Experts Yes Rate
Q1: Vignette is informative	0.88	0.92	0.84	0.76
Q2: Vignette is self consistent	0.72	0.88	0.80	0.68
Q3: Vignette is realistic	0.88	0.76	0.96	0.68
Q4: Simulator portrays correct communication style	0.88	0.88	1.00	0.76
Q5: Simulator is consistent with vignette	0.96	0.96	1.00	0.92
Q6: Simulator demonstrates barrier	1.00	1.00	1.00	1.00
Q7: Coach diagnoses reasonable barrier	0.96	1.00	0.92	0.92

M Diabetes Baseline vs. Test Comparison Questions

Human experts were asked the following evaluation questions when comparing baseline vs. test set conversations. As previously described, the baseline is conversations generated using synthetic users from demographics-only data, while test synthetic users had the same demographics data, plus the rich health data and the COM-B barrier.

1. Which of the two portrays one consistent barrier from the patient? (Y / N / Similar)
2. Which of the two conveys the original barrier more accurately? (Y / N / Similar)
3. Which of the two is more informative in understanding the patient? (Y / N / Similar)

N Reference Distribution vs. LLM-Generated Distribution of Barriers in Simulated Conversations

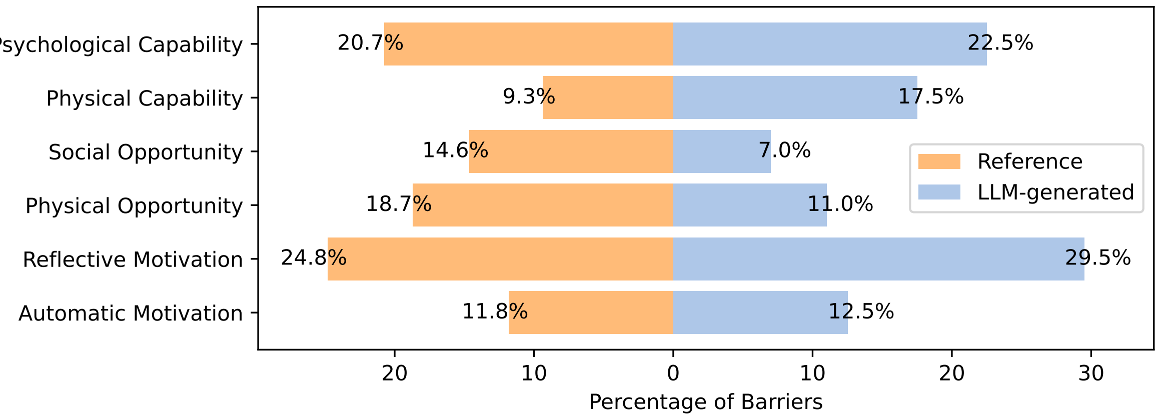

Figure A1: **Reference distribution vs LLM-generated distribution of barriers observed in simulated conversations**. The reference is the distribution of barriers observed in cardiometabolic behavioral literature. The LLM-generated distribution are extracted by the diabetes coaching agent in 200 simulated conversations.

O Comparison With Gemma 2 IT 27B Model

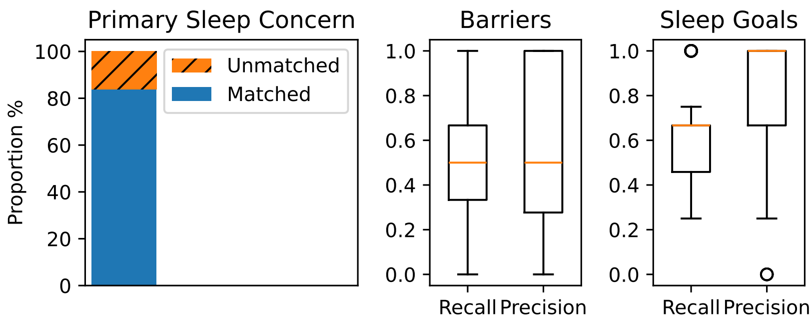

Figure A2: **Extracting synthetic users' sleep profile from interaction for automated evaluation of coaching agent, using Gemma 2 IT 27B instead of Gemini 1.5 Pro for synthetic user**. Our synthetic users (N=68) with sleep conditions interacted with our sleep agent for 10 turns. Recall and precision were computed by comparing the internal state-of-user model in the coaching agent to the true sleep profile of the synthetic user. "Primary sleep concern" is a single natural language statement, while "barriers" and "sleep goals" include multiple items, all in natural language. cf. Figure 2.

The following table includes a comparison of the key auto-evaluation metrics between synthetic users based on Gemma 2 IT 27B and Gemini 1.5 Pro.

Metric	**Gemma 2 IT 27B**	**Gemini 1.5 Pro**
Primary sleep concern accuracy %	83.8	89.7
Barriers recall (mean) %	45.2	71.4
Barriers precision (mean) %	55.5	72.5
Sleep goals recall (mean) %	61.3	66.4
Sleep goals precision (mean) %	82.6	84.2

P Error Analysis and Failure Modes

We manually inspected the 10.3% (7 out of 68) of the synthetic users from LifeSnaps for which our sleep coaching agent incorrectly determined their primary sleep concern. Qualitatively those 7 examples could be divided into two categories. One category was when the true primary concern and the coaching agent's prediction of the primary concern were related, but they did not pass the bar for paraphrase detection, such as "maintaining consistent high-quality sleep" vs. "optimizing already good sleep". The other category was when the synthetic users communicated multiple other concerns in addition to their primary sleep concern, and the coaching agent incorrectly determined one of the additional concerns as primary concerns, e.g. "weight management and its impact on sleep quality" vs. "difficulty falling asleep due to stress, leading to low energy and impacting social life." We did not observe a complete failure of synthetic users in communicating their given issues.

Q Agent Architecture Details

Our sleep coaching agent followed the agent architecture in Christakopoulou et al. (2024), which utilizes a novel Talker-Reasoner architecture and an explicit, structured model of the user. For our diabetes coaching agent, we used the agent developed by Yang et al. (2024), that probes for and identifies the root cause of an individual's dietary struggles (barriers) and delivers tailored tactics to overcome the identified barriers.

R Synthetic User-Sleep Coach Interaction Example

The following is an example interactions between the sleep coaching agent and two synthetic users we studied, the "full" synthetic user and a demographics-only synthetic user (Section 4.3). It is also an example of the expert annotation questions we used to generate Figure 3. For this particular question, the "full" synthetic user was option A, while the demographics-only synthetic user was option B.

> Question: The following is a description of a hypothetical person, who is interacting with a sleep coaching agent. Based on this profile, which one of interaction A and interaction B more closely matches the interaction between this person (=USER) and the agent (=COACH)?
>
> Description: "William is male. He is less than 30 years old. He typically sleeps for 9 hours at night. His sleep duration is relatively consistent. His sleep efficiency is typically at 93%. He is underweight. He is introverted. He is not agreeable. He is highly conscientious. He feels highly stable. He is moderately intellectual."
>
> Note: What the COACH says is shown here to provide conversational context, but please ONLY use what USER says to determine your answer.
>
> Answer: A / B ?
>
> ===== Interaction A =====
>
> [USER] I'm concerned about maintaining consistent sleep quality and duration, especially with increasing work demands. I'd like to explore strategies for maintaining a 9-hour sleep schedule and optimizing my sleep environment.
>
> [COACH] Hi there! It's great that you're prioritizing your sleep, especially with a busy work life. Maintaining consistent sleep can be a real game-changer! So, what feels like the *biggest* hurdle for you in getting those 9 hours of quality sleep?
>
> [USER] It's a combination of things. I find it hard to say no to extra work, which often leads to late nights. Also, I'm sensitive to noise and temperature changes, so creating the ideal sleep environment is a challenge.
>
> [COACH] So, juggling work and creating a peaceful sleep sanctuary are your main concerns. It sounds like saying "yes" to extra work sometimes means saying "no" to sleep. Can you tell me a bit more about your typical bedtime routine?
>
> [USER] I usually aim for a 10:30 PM bedtime to get my 9 hours, but it often ends up being closer to 11:00 PM or even later due to work. I try to keep my room cool and dark, but I live in a noisy apartment building, which can be disruptive.
>
> [COACH] That 10:30 PM bedtime goal is excellent! It shows you're really committed to those 9 hours. It sounds though, like work often pushes that back. And the noisy apartment… yikes! Do you have any pre-sleep rituals that help you wind down, even when work runs late?
>
> [USER] I don't have a consistent wind-down routine, which might be part of the problem. Sometimes I read before bed, but often I just collapse into bed exhausted after work.
>
> [COACH] So, sometimes reading, sometimes collapsing...it sounds like having a more consistent wind-down routine might be something to explore. How would a good night's sleep improve your work life, do you think?

[USER] A good night's sleep would definitely improve my focus and productivity at work, allowing me to handle those demanding tasks more efficiently. It would also improve my overall mood and energy levels, making me more pleasant to be around, even though I'm naturally introverted.

[COACH] More focus, better productivity, improved mood... those are *fantastic* benefits! And hey, being more pleasant to be around is a win-win for everyone, introvert or not! If we could wave a magic wand and give you the *perfect* sleep, what would that look like for you?

[USER] The perfect sleep would be consistently getting those uninterrupted 9 hours in a quiet, cool, and dark room, regardless of my workload. Waking up feeling refreshed and energized every morning would be ideal.

[COACH] Consistently getting 9 uninterrupted hours in a cool, dark, and quiet room, waking up refreshed... I love it! That's a great vision! So, if we were to focus on just *one* thing to move you closer to that vision this week, what would it be?

[USER] I think focusing on establishing a consistent wind-down routine would be the most impactful first step. Perhaps something simple to start, like a short meditation or reading for a set amount of time before bed, even if work runs late.

[COACH] That's a fantastic starting point! A consistent wind-down routine can make a world of difference. Short meditations or reading are both great options. Even 10-15 minutes can signal to your body that it's time to sleep. How about we aim for 15 minutes of either meditation or reading before bed every night this week? We can also talk about strategies for dealing with the noise in your apartment, and tweaking your sleep environment to make it even more conducive to rest. How does that sound?

[USER] That sounds like a manageable commitment. A consistent 15-minute wind-down routine before bed, even if it's just reading, seems like a good first step towards improving my sleep.

[COACH] Great! I'm glad that feels doable. Reading for 15 minutes is a wonderful way to signal to your body that it's time to rest. To keep a consistent body clock, try reserving the hour before bed for activities that help you unwind. Simple, hands-on tasks like tidying up, doing dishes, or preparing your room for sleep are an easy and effective way to wind down after a long day. Repeating these simple tasks each night builds a routine that signals to your body when it's time to sleep. We'll check in next week and see how it goes, and then we can talk about some strategies for creating a quieter sleep space.

[USER] That sounds good. I'll try the 15-minute reading routine before bed and reserving the hour before for winding down activities. I appreciate the suggestions.

[COACH] You've got this! Remember, small, consistent steps lead to big changes. I'm here to support you every step of the way. Have a great week, and I look forward to hearing how it goes!

[USER] Thanks. I'll let you know how the reading goes and if I have any trouble quieting down my apartment.

[COACH] Sounds good! Remember, consistency is key. Even if some nights don't go perfectly, just keep coming back to that 15-minute wind-down routine. I'm looking forward to hearing about your progress.

===== Interaction B =====

[USER] I'm struggling with poor sleep quality, it's affecting my work. I'd like to fall asleep faster, reduce nighttime awakenings, and wake up feeling refreshed so I can perform better at work, improve my mood, and have more energy for my hobbies, but I'm stressed about my work/life balance, have an inconsistent sleep schedule, and use electronics before bed.

[COACH] It sounds like you're juggling a lot! It's totally understandable that you're struggling with sleep when work, mood, hobbies, and everything in-between is on your mind. So, out of all of those things, what feels like the *biggest* sleep challenge right now?

[USER] I think the biggest challenge right now is my inconsistent sleep schedule because of social events and commitments. It makes it really hard to stick to a regular sleep routine, which I know is important for good sleep.

[COACH] It makes total sense that a wonky sleep schedule makes it tough to get good rest. Can you tell me a bit more about what a typical week's sleep schedule looks like for you? What time are you hitting the hay and waking up, even if it varies?

[USER] Sure. My bedtime and wake-up times vary quite a bit throughout the week. Weekdays, I usually try to be in bed by 11 PM, but with work events and social gatherings, it often ends up being closer to midnight or even later. Then I have to wake up around 7 AM for work, so I rarely get a full night's sleep.

[COACH] So if I'm hearing you right, weekdays are a bit of a sleep rollercoaster thanks to work and social stuff - sometimes bed by 11 p.m., sometimes midnight or later, but *always* up at 7 a.m. What about weekends? Is it any easier to stick to a schedule then?

[USER] Weekends are a different story. I usually sleep in until 9 or 10 AM, sometimes even later if I've been out late the night before. Then I stay up late, often past midnight, so my sleep schedule gets completely flipped.

[COACH] So on weekends, you're catching up on sleep by sleeping in, but then staying up late, which totally throws off your weekday rhythm. It sounds like you're dealing with "social jet lag," where your sleep schedule does a 180 on the weekends. How do you think this shift affects you during the week?

[USER] It makes it really hard to get back into a regular routine on Sunday night, so I'm often tired and sluggish on Mondays, which impacts my work performance.

[COACH] Mondays sound like a rough start to the week thanks to that weekend sleep shift impacting your work. If you *could* wave a magic wand and get your sleep sorted, what would the *perfect* sleep scenario look like for you?

[USER] Ideally, I'd like to fall asleep around 10:30 PM every night and wake up feeling refreshed at 7 AM, even on weekends. That way, I'd get a consistent 8-8.5 hours of sleep, which I think would make a huge difference in my energy levels and overall well-being.

[COACH] So, the dream is 10:30 p.m. to 7 a.m. - consistent, even on weekends - giving you that solid 8 to 8.5 hours of ZZZs. I love it! Do you have any ideas about how you might start moving towards that ideal bedtime and wake time, even if it seems difficult right now?

[USER] I think I could start by setting a consistent bedtime alarm, even on weekends, to remind myself to start winding down. Maybe even try shifting my bedtime 15 minutes earlier each night until I reach that 10:30 PM goal.

[COACH] That's a great idea! Setting a consistent bedtime alarm, even on weekends, can really help regulate your body's natural sleep-wake cycle. Shifting your bedtime 15 minutes earlier each night is a smart approach, too – it's gradual, so your body can adjust more easily.

While setting an alarm for *bedtime* is a fantastic first step, expert clinicians often suggest a slightly different approach to figuring out your *ideal* wake-up time. They recommend trying to wake up *without* an alarm for a week or two, even on weekends if possible, and noting down when your body naturally wakes up. This helps your body catch up on any sleep debt you might have, and it'll give you a really good sense of how much sleep you truly need each night. This natural wake-up time becomes your anchor point, and from there, you can count back 8 to 8.5 hours to find your ideal bedtime.

Once you have a better understanding of your natural sleep needs, you can then fine-tune your bedtime routine and create a relaxing wind-down ritual to help signal to your body that it's time to sleep. How does that sound? We can brainstorm some wind-down ideas together, too, if you'd like.

[USER] That sounds like a good approach. I'm willing to try waking up without an alarm to see what my natural wake-up time is, and then work backward from there to figure out my ideal bedtime. I'd also appreciate some brainstorming on wind-down routines, as that's something I've struggled with.

[COACH] Great! I'm happy to hear you're open to trying this. Let's talk wind-down routines. A good wind-down routine is like telling your body, "Okay, it's time to shift gears and get ready for sleep." Think of it as creating a relaxing and predictable pre-sleep ritual that you can look forward to. Here are a few ideas we can explore, and we can tailor them to what you enjoy and what works for *you*:

* **Dim the lights:** Lowering the light levels in your home about an hour or two before bed helps your body start producing melatonin, the sleep hormone. Think soft lamps, candles (safely!), or even dimmer switches. * **Screen-free time:** The blue light emitted from electronic devices like phones, tablets, and computers can interfere with melatonin production and make it harder to fall asleep. Aim to put your screens away at least an hour before bed. * **Warm bath or shower:** A warm bath or shower can be incredibly relaxing. The rise and subsequent fall in body temperature after getting out of the water can actually help promote sleepiness. * **Relaxing activities:** Think about activities that help calm your mind and body. Reading a physical book (not an e-reader!), listening to calming music, gentle stretching, or journaling are great options. * **Mindfulness and meditation:** Even a few minutes of mindfulness or meditation can help quiet a racing mind and reduce stress. There are plenty of guided meditation apps available if you're new to this. * **Consistent sleep environment:** Make sure your bedroom is dark, quiet, and cool. A comfortable temperature, a good mattress, and cozy bedding can make a world of difference.

What sounds most appealing to you from this list? Or are there other things you've tried in the past, or are curious about trying now, that we should add to the list?

[USER] I like the idea of a warm bath or shower before bed, and I definitely need to work on the screen-free time. I'm also curious about trying some mindfulness exercises, maybe a short guided meditation to help me relax.

[COACH] Those are all fantastic choices! A warm bath or shower, screen-free time, and mindfulness exercises can be powerful tools for a relaxing wind-down routine.

Let's break this down into smaller, manageable steps:

1. **Screen-Free Time:** Since you mentioned this is a big one, how about starting with just 30 minutes of screen-free time before bed? You can gradually increase this time as you get more comfortable. Perhaps set an alarm on your phone as a reminder, and then put your phone in another room—out of sight, out of mind!

2. **Warm Bath/Shower:** Is a bath or shower more appealing to you? Schedule it into your evening routine. Perhaps it could be right after your screen-free time begins. Adding some Epsom salts or essential oils like lavender to your bath can enhance the relaxation factor even more.

3. **Mindfulness/Meditation:** Start with short, guided meditations, even just 5-10 minutes. There are many free apps and online resources available. Insight Timer, Calm, and Headspace are just a few examples. Experiment and find one that you connect with.

So a sample routine could be: Finish any necessary screen activities, set an alarm for 30 minutes before your target bedtime, then take your warm bath/shower. Once you are out, try a short guided meditation. Once you start to feel drowsy, head to bed. Continue this until it feels like a normal routine. How does this sound as a starting point? We can always adjust it as needed.

[USER] That sounds like a great starting point. I think I can manage 30 minutes of screen-free time before bed, followed by a warm bath and a short meditation. I prefer baths, so I'll try adding some Epsom salts or lavender oil for extra relaxation.

[COACH] Wonderful! It's great that you're starting with what feels manageable and enjoyable. Thirty minutes of screen-free time, a warm bath with Epsom salts or lavender oil, and a short meditation sounds like a truly relaxing pre-sleep ritual. Remember, consistency is key. Try to stick to this routine as much as possible, even on weekends, to help regulate your sleep-wake cycle.

Let's check in again next week. How about we schedule a follow-up for then to see how the new routine is going, discuss any challenges you might be experiencing, and tweak things as needed? We can also discuss how the alarm-free wake-up experiment is going and start figuring out that ideal bedtime based on your natural wake time.