# Towards a Personal Health Large Language Model

Justin Cosentino*, Anastasiya Belyaeva*§, Xin Liu*§, Nicholas A. Furlotte*, Zhun Yang‡, Chace Lee‡, Erik Schenck‡, Yojan Patel‡, Jian Cui‡, Logan Douglas Schneider‡, Robby Bryant , Ryan G. Gomes , Allen Jiang , Roy Lee , Yun Liu , Javier Perez , Jameson K. Rogers , Cathy Speed , Shyam Tailor , Megan Walker , Jeffrey Yu , Tim Althoff , Conor Heneghan , John Hernandez , Mark Malhotra , Leor Stern , Yossi Matias , Greg S. Corrado , Shwetak Patel , Shravya Shetty , Jiening Zhan , Shruthi Prabhakara , Daniel McDuff†§, and Cory Y. McLean†§

Google LLC
*Co-first author.
‡Core contributor.
†Co-last author.
§Corresponding authors: {belyaeva,xliucs,dmcduff,cym}@google.com.

## Abstract

In health, most large language model (LLM) research has focused on clinical tasks. However, mobile and wearable devices, which are rarely integrated into such tasks, provide rich, longitudinal data for personal health monitoring. Here we present Personal Health Large Language Model (PH-LLM), fine-tuned from Gemini for understanding and reasoning over numerical time-series personal health data. We created and curated three datasets that test 1) production of personalized insights and recommendations from sleep patterns, physical activity, and physiological responses, 2) expert domain knowledge, and 3) prediction of self-reported sleep outcomes. For the first task we designed 857 case studies in collaboration with domain experts to assess real-world scenarios in sleep and fitness. Through comprehensive evaluation of domain-specific rubrics, we observed that Gemini Ultra 1.0 and PH-LLM are not statistically different from expert performance in fitness and, while experts remain superior for sleep, fine-tuning PH-LLM provided significant improvements in using relevant domain knowledge and personalizing information for sleep insights. We evaluated PH-LLM domain knowledge using multiple choice sleep medicine and fitness examinations. PH-LLM achieved 79% on sleep and 88% on fitness, exceeding average scores from a sample of human experts. Finally, we trained PH-LLM to predict self-reported sleep quality outcomes from textual and multimodal encoding representations of wearable data, and demonstrate that multimodal encoding is required to match performance of specialized discriminative models. Though further development and evaluation are necessary in this safety-critical domain, these results demonstrate the broad knowledge and capabilities of Gemini models and the benefit of contextualizing physiological data for personal health applications.

## 1 Introduction

Large language models (LLMs) are versatile tools for generating language and have shown strong performance across a range of diverse domains. LLMs achieved passing grades on the US legal bar exam [30] and second year medical school exams [45, 50, 55]. In medicine in particular, natural language as an interface has shown potential to influence clinical practice [37], education, and research [39]. When enriched with healthcare data, LLMs attain impressive performance in medical question-answering [50, 55], analysis of electronic health records [63], differential diagnosis from medical images [58], assessment of psychiatric functioning based on standardized assessments [20], and psychological intervention delivery [34, 52, 53]. Strong performance on these tasks shows that LLMs are able to effectively capture signal from "clinical data" collected within a clinical setting.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

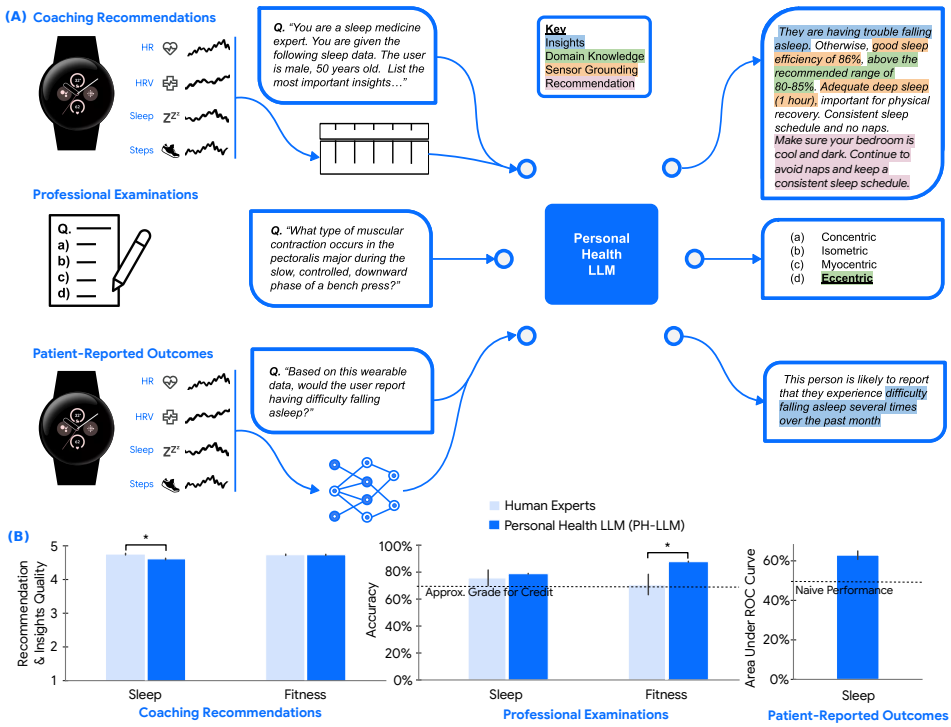



Figure 1: **PH-LLM: A Personal Health Large Language Model.** **(A)** We present PH-LLM, a version of Gemini fine-tuned for personal health and wellness. We evaluated PH-LLM on three aspects of personal health: generating personalized insights and recommendations for user goals in the domains of sleep and fitness, assessing levels of expert knowledge from certification examination style multiple choice questions, and predicting patient-reported outcomes in sleep quality from detailed sensor information. **(B)** Performance of PH-LLM contextualized with expert human responses. Error bars represent 95% confidence intervals. "$*$" indicates a statistically significant difference between two response types. "Naive Performance" is that achieved by a random classifier. Human expert performance is not available for patient-reported outcome prediction from sensor features as this is not commonly performed, and no fitness-related outcomes were measured in the study [36].

Due to their sporadic nature, conventional clinical visits often fail to capture key aspects of human health and well-being that are measurable with wearable devices including sleep, physical activity, stress, and cardiometabolic health measured through physiological response and behavior. These continuous, longitudinal measures have significant advantages for health monitoring in that they are passively and continuously acquired, and provide direct physiological and behavioral signals. However, they have neither been deeply integrated into clinical practice nor incorporated into standard datasets used for medical question-answering [23, 29], despite statistics on adverse health outcomes, morbidity, and Disability-Adjusted Life Years that underscore the profound impact these factors have on overall health [3, 11, 18, 46, 60]. This limited adoption is likely because these data are typically captured without context, are computationally demanding to store and analyze, and can be difficult to interpret. As a result, general foundation LLMs or even medically-tuned LLMs may lack the ability to use these data to reason about and recommend interventions based on individual health behaviors.

Mobile devices, including smart wearables and smartphones, have become instrumental tools for monitoring personal health metrics and gathering longitudinal data that cannot be obtained in traditional clinical settings [56]. Unlike structured clinical data, personal health data is heterogeneous across data types, sources, and timescales [36], ranging from continuous streams of biometric data from wearables, such as sleep patterns or heart rate, to sporadic and qualitative inputs like exercise logs, dietary logs, mood journals, and even social media activity.

In this paper, we introduce Personal Health Large Language Model (PH-LLM), a version of Gemini fine-tuned to generate both insights about and recommendations to improve personal health behaviors related to sleep and fitness patterns. We evaluate the performance of PH-LLM across three tasks: coaching recommendations, multiple choice exams assessing expert knowledge, and prediction of subjective patient-reported outcomes (PROs). The coaching recommendations tasks are tailored to two

verticals of high personal health interest: sleep and fitness. The sleep tasks leverage individual's sleep metrics to derive insights, possible etiological factors, and provide personalized recommendations to improve sleep quality. The fitness tasks integrate information from training load, sleep, health metrics, and subjective feedback to provide personalized recommendations for the intensity of a physical activity that day. For the coaching recommendations tasks, we create the first personal health case study dataset to benchmark LLM performance on reasoning and understanding of personal health behaviors. This dataset consists of long-form questions grounded in summarized personal health behavior data, vertical-specific evaluation rubrics, and expert human responses for 857 case studies across sleep and fitness. Through rigorous human and automatic evaluation, we demonstrate that Gemini Ultra 1.0 approaches expert performance in fitness while fine-tuning PH-LLM enables it reduce the gap with experts in sleep coaching experiences, explore the breadth of personal health knowledge encoded within Gemini models, and apply multimodal capabilities to prediction of PROs in sleep (Figure 1). Our key contributions are summarized as follows:

- **PH-LLM:** We introduce a new model fine-tuned from Gemini for applications in personal health, capable of performing interpretation of time-series sensor data from wearables (i.e., Fitbit and Pixel Watch) for analysis and recommendations in sleep and fitness.
- **Long-form case studies from domain experts:** We create the first dataset of detailed personal health case studies in sleep and fitness curated by multiple experts in the associated domains. The dataset contains individual wearable sensor data and corresponding long-form insights and recommendations. We present rubrics for evaluation of long-form responses that span domain knowledge, use of user data, personalization, and potential for harm, and provide insights on training experts for accurate evaluation.
- **Benchmark and contextualize personal health question-answering:** We curate a set of validated domain-specific multiple choice examination questions on sleep and fitness, establish strong benchmarks based on continuing medical education requirements, and provide context for the scores through a set of human experts who completed a representative set of exam questions.
- **Multimodal sensor interpretation of self-reported outcomes:** We successfully integrate longitudinal time-series sensor features to interpret a user's subjective experience. To do so, we evaluate the capabilities of PH-LLM in predicting sleep disturbance and impairment PROs (acquired through validated survey instruments) from passive sensor readouts and show that accurate model performance requires native multimodal data integration.

## 2 Personal Health Dataset Creation and Methods

Owing to the absence of clearly defined language and multimodal datasets in the domain of personal health, we created datasets and associated tasks to evaluate different capabilities of PH-LLM. These datasets include case studies about real-world coaching recommendations, professional examinations that test domain knowledge about sleep medicine and fitness, and patient-reported sleep outcomes.

### 2.1 Coaching recommendations

Many real-world applications of LLMs for personal health require realistic long-form text generation, which is challenging to evaluate automatically. As previously observed in the medical domain, strong performance on question-answering tasks does not necessarily transfer to the complexity of real-world tasks [17]. To address the absence of rich long-form tasks for personal health data, in conjunction with domain experts and overseen by clinical leads we created detailed case studies that span two key personal health domains: sleep and fitness. Each case study was designed to interpret a range of physiological sensor information toward deriving insights, potential causes, or recommendations for future behaviors, and was sampled from high-volume anonymized production data from individuals who provided consent for research purposes.

The sleep case studies aimed to enhance understanding of sleep patterns, identify causes of irregular sleep, and offer actionable recommendations based on these findings. Each case study incorporated wearable sensor data for up to 29 days, demographic information, and an expert analysis (Figure 2A-C). This comprehensive approach both facilitates a deeper understanding of health-related behaviors and also guides the development of personalized interventions to improve individual outcomes.

The fitness case studies were designed to provide a comprehensive analysis of an individual's training load, sleep patterns, and health metrics, and were similarly based on wearable sensor data over 30

days, demographic information, and expert analysis (Figure 2D-G). The goal of the fitness-related task was to synthesize the suite of metrics into a data-driven assessment of the extent to which the individual is prepared for physical activity today and provide associated recommendations.

**Sleep Case Study Creation.** In the development of sleep case studies, we recruited six domain experts in sleep medicine to craft guidance in the second person narrative, fostering a direct and personalized dialogue with the user. The six sleep experts all possessed advanced degrees (M.D., D.O., or Psy.D.) in sleep medicine and professional sleep medicine work experience ranging from 4 to 46 years. All experts were trained to read and interpret wearable data and map outputs to their corresponding sleep medicine literature counterparts. Experts were instructed to use best practices in goal-setting, emphasizing the creation of recommendations that are Specific, Measurable, Achievable, Relevant, and Time-bound (SMART) [15]. The data was sampled to achieve a representative group across age and gender (see Appendix D.1.1 for details, Figure D.2).

As illustrated in Figure 2A-C, demographics information (age and gender), daily sleep metrics (e.g., bedtimes, wake times, and sleep stage durations), and aggregated sleep statistics (e.g., average bedtime) were selected collaboratively with the experts (Appendix D.1.1). These data were presented to the experts in both graphical and tabular formats for ease of analysis (Figure 2A-C, Table D.4). Finally, the experts were tasked with composing responses across three sections (insights, etiology,

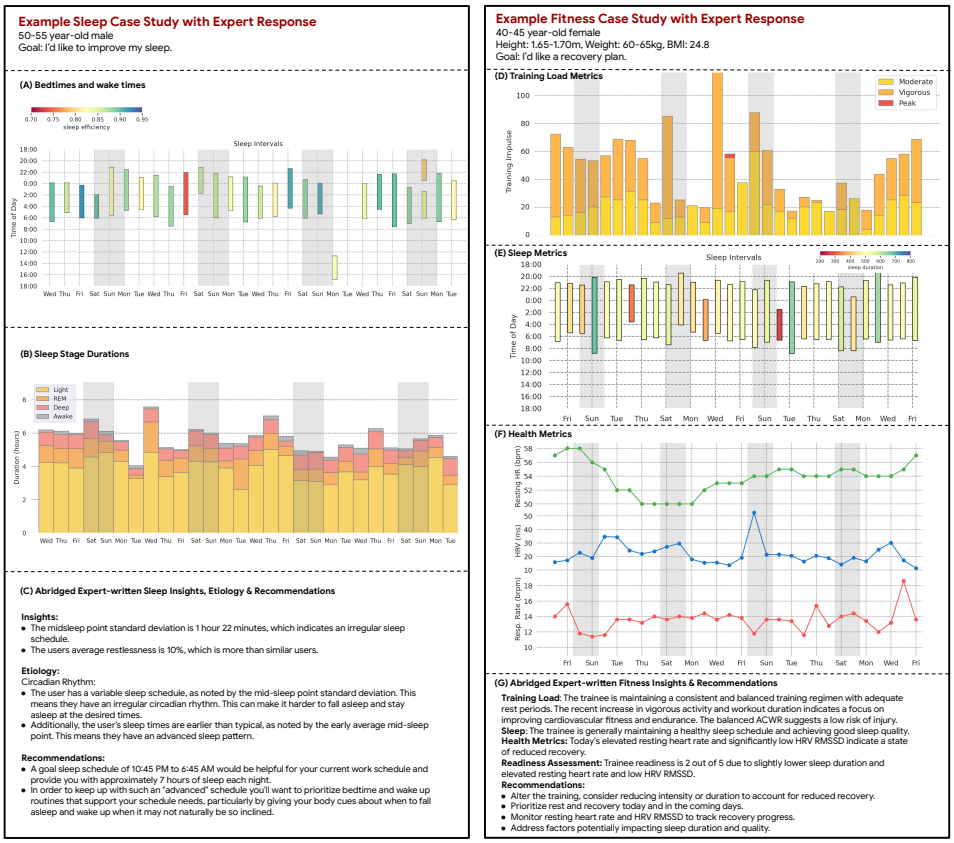



Figure 2: **Case study examples**: wearable sensor data inputs and expert analysis and recommendations for **(A-C)** sleep and **(D-G)** fitness. For sleep, experts considered an individual's demographics and sensor data for up to 29 days including daily metrics of **(A)** bedtimes and wake times and **(B)** time spent in various sleep stages and awake (see Table D.4 for all metrics). Experts also analyzed aggregated statistics of sleep metrics (Table D.5). Experts composed responses **(C)** including insights, potential etiology, and recommendations for improving sleep quality. For fitness, experts considered an individual's demographics and sensor data over a 30-day period including daily metrics of **(D)** cardiovascular training load, **(E)** sleep metrics, and **(F)** health metrics such as resting heart rate, heart rate variation, and respiratory rate (Tables D.11-D.17). Experts composed responses **(G)** including insights and provided a workout readiness assessment and recommendations.

and recommendations), aimed at analyzing the data with the objective of enhancing the sleep quality of the individual under consideration. For details on each section, see Appendix D.1.1.

**Fitness Case Study Creation.** To construct fitness case studies (Figure 2D-G), we recruited seven domain experts in fitness to analyze an individual's quantitative fitness data. The seven fitness experts all possessed advanced degrees (M.S., M.A., M.Ed., or D.A.T.) related to the athletic training field and professional athletic training work experience ranging from 4 to 25 years. The experts were directed to formulate insights, assessments, and recommendations in the second person narrative. The data for fitness case studies were sampled to produce a variety of different fitness assessments (see Appendix D.1.2 for details). The quantitative fitness data included a comprehensive array of metrics encompassing daily cardiovascular training load, sleep patterns, and health metrics spanning the preceding 30-day period (see Appendix D.1.2 for details). These data were presented in tabular, text, and graphical formats. The experts were tasked with providing responses to four sections (training load, sleep, health metrics, and assessment), with the objective of facilitating a personalized approach to improving individual fitness levels by guiding on the intensity and duration of fitness sessions.

**Holistic View of Case Study Creation.** For both the sleep and fitness verticals, we generated two sets of data: a dataset used for model training, validation, and testing and (457 case studies for sleep, 300 for fitness) a holdout dataset (50 case studies for sleep, 50 for fitness) that was only used for final evaluation of the model by experts (Appendix D.1.3, Figure D.1).

### 2.2 Professional examinations

**Sleep Medicine Exams.** We compiled a set of 629 multiple choice questions (MCQs) from BoardVitals [8] sleep medicine board review question banks. We used text exam questions from the American Medical Association (AMA) Physician's Recognition Award (PRA) "Category 1 - Sleep Medicine" question bank, which emulates exam content for the American Board of Internal Medicine (ABIM) Sleep Medicine Certification Exam. We also used text exam questions from the Sleep Medicine Maintenance of Certification (MOC) Exam and Longitudinal Knowledge Assessment Review question bank, which emulates exam content for the ABIM Sleep Medicine MOC Exam and ABIM Longitudinal Knowledge Assessment. This compiled set of MCQs spanned a wide range of sleep-related topics: Normal Sleep and Variants (N=127), Breathing Disorders (N=84), Hypersomnolence (N=60), Insomnias (N=85), Movement Disorders (N=23), Parasomnias (N=57), Sleep in Other Disorders (N=112), and Sleep-Wake Timing (N=81).

**Fitness Exams.** We compiled 99 multiple choice questions sourced from question banks that emulate exam content for the Certified Strength and Conditioning Specialists (CSCS) exam preparation book provided by the National Strength and Conditioning Association (NSCA) [42]. We used the test exam questions from the NSCA-CSCS textbook "Essentials of Strength Training and Conditioning".

Accuracy was used as the metric to evaluate the performance of our model in professional exams, in line with prior work evaluating MedMCQA [54]. Each exam question presents up to five possible answers, with a single correct answer, facilitating automated and quantitative assessment of performance. We did not train models directly on MCQs and all samples were used in evaluation.

### 2.3 Patient-reported outcomes

To evaluate the ability of PH-LLM to predict patient-reported outcomes (PROs) from longitudinal passive sensor data, we used a large IRB-approved study in which wearable data was collected for a population of 4,759 consented individuals for a four-week period [36]. At both intake and completion, participants completed the Patient-Reported Outcomes Measurement Information System (PROMIS) [41] short-form Sleep Disruption and Sleep Impairment surveys [67]. Both surveys contained eight items with answers on a 5-point Likert scale (Appendix F.3). The study thus linked individuals' perceived sleep quality and its impact on their functioning with longitudinal observed physiological (e.g., heart rate, sleep duration) and behavioral (activity) measurements.

To maximize sample size, we used the intake survey responses as the basis for prediction. For each question, we defined a binary outcome that compared the highest answer (e.g., "strongly agree") against all others (Figure F.2). Features used to predict each binary outcome included 20 time-varying wearable measurements (Table F.1), each of which was collected from study participants over a four-week span. After filtering and missing value imputation (see Appendix F.1), we obtained 4,978 training examples, 703 validation examples and 1,433 examples.

### 2.4 Methods

To train PH-LLM, we fine-tuned Gemini Ultra 1.0 on coaching recommendations and additionally trained an MLP adapter to encode multimodal wearable measurements collected as part of the patient-reported outcomes dataset. We then evaluated PH-LLM on the dataset of coaching recommendations by asking domain experts to grade the responses across variety of dimensions using a 5-point Likert scale as well as with automatic evaluation. We computed accuracy for the professional examinations dataset, and AUROC and AUPRC for patient-reported outcomes. For details, see Appendix C.

## 3 Results

### 3.1 PH-LLM approaches expert performance on long-form case studies

We evaluated the aggregated performance of PH-LLM and human experts on the long-form case study responses, rated by human experts using 15 questions with grading scale 1 through 5, spanning topics such as using important domain knowledge, correctly referencing relevant user data, and avoiding confabulations. A rating of 5 indicates high quality: a 2 or 3 indicates many or several important data interpretations are missing, while a 4 or 5 indicates few or none missing. All 15 questions and rating descriptions are detailed in Appendix D.2. For sleep case studies, PH-LLM received an average rating of 4.61 versus 4.75 for human experts, indicating a close match ($p = 3.3 \times 10^{-11}$, $N \geq 2606$, $Z = -6.63$ $r = -0.08$, Figure 1A). Although the difference is statistically significant, the effect size is small and our model responses are high quality as indicated by receiving the top rating of five 73% of the time. Fine-tuning PH-LLM on sleep case studies significantly improved its overall performance in this task (average rating of 4.51 versus 4.61, $p = 4.0 \times 10^{-6}$, $N \geq 2603$, $Z = 4.63$, $r = 0.06$). For fitness case studies, PH-LLM aggregate performance was not statistically different from expert performance ($p = 0.48$, $N \geq 3335$, $Z = -0.70$, $r = -0.01$, Figure 1B). Gemini Ultra 1.0 responses were also statistically indistinguishable from human expert responses ($p = 0.92$, $N \geq 3161$, $Z = -0.10$, $r = -0.00$). Furthermore, we conclude moderate inter-rater reliability, as Gwet's AC2 [26, 27] ranged from 0.699 to 0.956 (Appendix D.4).

Since the case studies consist of multiple sections, we also analyzed ratings for each section separately (Figure 3A). For sleep case studies, fine-tuning PH-LLM improved its ability to provide insights and etiologies ($p = 6.65 \times 10^{-7}$, $N \geq 800$, $Z = 5.18$, $r = 0.11$ and $p = 2.46 \times 10^{-3}$, $N \geq 801$, $Z = 3.15$, $r = 0.07$, respectively), with recommendations showing no statistically significant difference ($p = 0.45$, $N \geq 801$, $Z = 0.76$, $r = 0.02$). We further analyzed ratings by various rubric questions. Fine-tuning PH-LLM improved its ability specifically on being able to reference important domain knowledge ($p = 4.47 \times 10^{-5}$, $N \geq 201$, $Z = 4.44$, $r = 0.19$), important interpretations ($p = 4.47 \times 10^{-5}$, $N \geq 201$, $Z = 4.48$, $r = 0.19$), important user data ($p = 5.21 \times 10^{-8}$, $N \geq 201$, $Z = 5.91$, $r = 0.26$), and no unimportant interpretations ($p = 4.31 \times 10^{-2}$, $N \geq 201$, $Z = 2.53$, $r = 0.11$), see Figure D.3. Overall, these results suggest that fine-tuning improved the model's ability to mention relevant domain knowledge, relevant interpretations, and relevant user data, especially when deriving insights and etiology from the data.

For fitness case studies, PH-LLM had similar performance (no statistically significant difference detected, $N \geq 768$) to human experts on three out of four sections (Figure 3B). Training load was the only section in which responses from human experts were rated higher than those from PH-LLM ($p = 0.01$, $N \geq 768$, $Z = 3.02$, $r = 0.07$). When analyzing ratings by rubric questions, we observed no statistically significant differences in ratings between PH-LLM and human experts (Figure D.3).

Furthermore, our fine-tuned AutoEval models can act as strong proxies for expert annotation. The best AutoEval models ranked case study response sources similarly to human experts (compare Figure D.8 to Figure 3). When measuring Spearman's rank correlation, Kendall's Coefficient of Concordance (Kendall's W), and Weighted Cohen's Kappa between AutoEval rating predictions and ground-truth human ratings across validation datasets, the best AutoEval models obtained similar prediction-rating agreement compared to inter-rater agreement metrics (Appendix D.4). We explored different AutoEval training data mixtures and found that all mixtures produced models that significantly improved upon a Gemini Pro 1.0 rater not explicitly fine-tuned for AutoEval tasks (Table D.22).

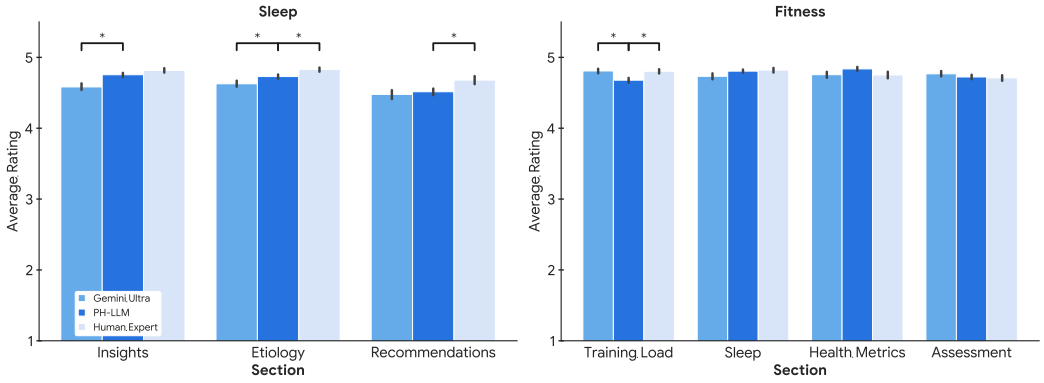



Figure 3: **Case Study Human Evaluation.** Mean expert ratings by subsections across the **(A)** sleep and **(B)** fitness domains. Error bars denote 95% confidence intervals. "$*$" indicates a statistically significant difference between two response types after multiple hypothesis testing correction.

### 3.2 PH-LLM exceeds grades to receive CME credit on sleep medicine and pass the fitness certification practice examination question banks

PH-LLM correctly answered 79% of sleep medicine and 88% of fitness board examination questions tested, comfortably exceeding the approximate grade (70%) to either receive CME credit for sleep or pass the practice exam for fitness (Table 1). On the AMA PRA Category 1 and ABIM MOC, PH-LLM scored 79% while Gemini Ultra 1.0 scored 77%. On the NSCA-CSCS coaching exams both PH-LLM and Gemini Ultra 1.0 scored 88%. The sleep medicine question bank contained additional metadata for each question including the distribution of responses from human test takers, enabling comparisons of performance by empirical question difficulty. Both PH-LLM and Gemini Ultra 1.0 performed comparably across the question difficulty strata and suggest that the performance of PH-LLM is comparable to that of humans who have prepared for or are in the process of preparing for these examinations (Table 2). To further contextualize the performance of PH-LLM with experts, five professional athletic trainers (average experience: 13.8 years) and five sleep medicine experts (average experience: 25 years) with advanced degrees were recruited to take the respective exams. The experts achieved an average score of 71% in the fitness exam and an average score of 76% in a representative subset of the sleep medicine exam (N=204) stratified based on medical content categories [4] and their difficulty levels. As illustrated in Table 1, PH-LLM outperforms expert graders on both professional exam question banks.

We performed ablation studies on the use of self-consistency [59] (N=5 rounds) and chain-of-thought (CoT) prompting [62]. Self-consistency improved performance on fitness questions for both CoT and Non-CoT prompting techniques while the performance from including CoT was mixed (Table E.1). See Appendix E for question prompts and Appendix E.2 for ablation results.

### 3.3 Multimodal sensor integration enables PH-LLM to predict patient-reported outcomes

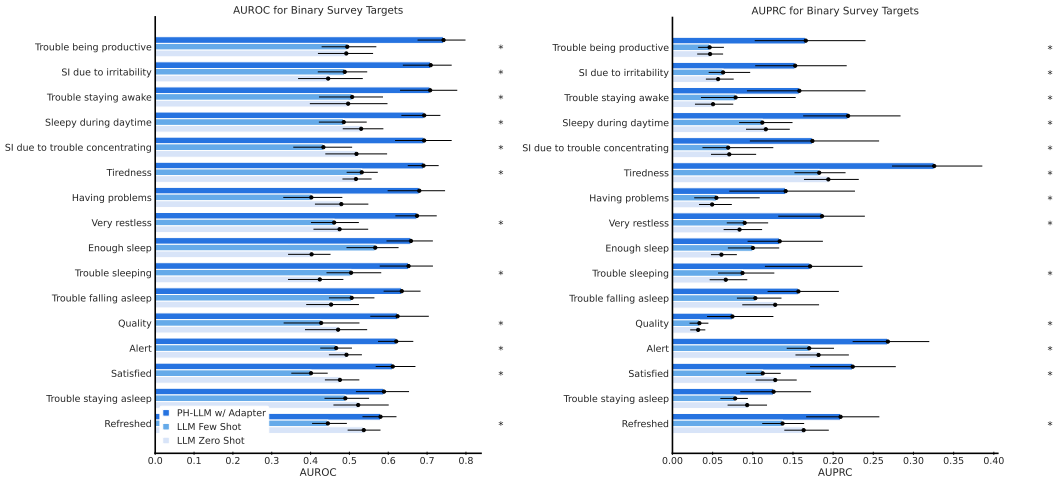
We evaluated the ability of PH-LLM to predict self-reported outcomes in sleep disturbance and sleep impairment. Using a dataset of 4,759 individuals with 20 wearable device measurements, a subset

Table 1: **Performance on Professional Exam Question Banks.** Accuracy on the multiple choice questions from AMA PRA Category 1 - Sleep Medicine and ABIM MOC - Sleep Medicine MOC and NSCA-CSCS coach certification examination question banks.

| **Sleep Medicine** | Approx. CME Grade | Expert | Gemini Ultra 1.0 | PH-LLM |
|---|---|---|---|---|
| AMA PRA Category 1 / ABIM MOC | 70%* | 76% | 77% | **79%** |
| **Fitness** | Approx. Pass Grade | Expert | Gemini Ultra 1.0 | PH-LLM |
| NSCA-CSCS Coaching Certification | 70%† | 71% | **88%** | **88%** |

* https://www.boardvitals.com/sleep-medicine-moc-recertification.

† https://www.nsca.com/certification/cscs/certified-strength-and-conditioning-specialist-exam-description




(A) (B)

Figure 4: **Prediction of Patient-Reported Outcomes.** Area under the **(A)** receiver operator characteristic and **(B)** precision-recall curves of PH-LLM, zero-shot, and few-shot prompting approaches when predicting binary survey response outcomes ( Appendix F.3). Error bars denote 95% confidence intervals and "*" denotes that PH-LLM w/ Adapter is significantly better than other models.

of the Google Digital Wellbeing Study dataset [36], and 16 derived binary outcomes, we trained a multilayer perceptron (MLP) adapter to map the 20 wearable measurements into PH-LLM's latent token space (Section C.4). We then provided the latent tokens to PH-LLM as context and prompted it to predict each patient outcome. Given the trained MLP adapter, we evaluated the likelihood of each binary outcome for each sample in the evaluation set and compared its predictive power to baseline approaches using in-context learning of textual sensor data representations.

We compared the area under the receiver operator characteristic curve (AUROC, Figure 4A) and area under the precision-recall curve (AUPRC, Figure 4B) for each binary trait, computed in the holdout set, for PH-LLM using the multimodal adapter and the zero- and few-shot text approches. We note that in general, objective measurements of sleep and sleep behaviors provide only modest predictive power for perceived sleep quality metrics. However, PH-LLM using the adapter significantly outperformed both prompt-based approaches in terms of both AUROC and AUPRC for 12 of the 16 traits (Tables F.18 and F.19). This relative performance increase is due to adapter-enabled LLMs being able to capture more signal from the training set as compared to zero and few-shot prompting which see a very limited amount of training data [7].

We do not expect an adapter-enabled language model to exceed the performance of a specialized discriminative model trained to predict the same binary traits. However, if the LLM has roughly comparable ability to a specialized model, this could be beneficial. To assess how well PH-LLM performed compared to a traditional machine learning approach, we fit a logistic regression (LR) model for each binary trait. Comparing PH-LLM to LR models trained using the same encoded vector input, we found no statistically significant differences in performance between for either AUROC or AUPRC (Tables F.3a, F.3b, F.18, and F.19).

Table 2: **Performance Comparison of Models and Experts Relative to Average Reported Test Takers for the Sleep Professional Exam.** Questions were classified as "Easy", "Medium", or "Hard" based on the percentage range of human test takers who answered the questions correctly.

| Difficulty | Count | Expert | Gemini Ultra 1.0 | PH-LLM |
|---|---|---|---|---|
| Easy (90%-100%) | 214 | 90% | 94% | **95%** |
| Medium (75%-90%) | 204 | **81%** | 78% | 80% |
| Hard (0%-75%) | 211 | 53% | 55% | **57%** |

## 4 Related Work

**Large language models in health** LLMs have the ability to perform complex language comprehension and reasoning tasks, generate coherent text and thereby enable real-world applications [5, 21, 47, 51, 57]. Explorations of LLM utility in health domains have shown their ability to answer medical questions and enable data-driven decision making [24, 44, 50, 54, 55, 66]. Med-PaLM [54] and its successor, Med-PaLM 2 [55], leveraged a combination of methodological advancements and domain-specific fine-tuning to increase performance, relative to previous models, on medically relevant evaluation tasks. Med-PaLM 2 achieved a score of up to 86.5% across several medical datasets, such as MedMCQA, PubMedQA, and MMLU clinical topics, achieving physician-level performance. GPT-4 and Gemini have further improved performance on the USMLE-style examinations in MedQA, reaching 90.2% [45] and 91.1% [50], respectively. On complex diagnostic tasks it is even possible for LLMs to outperform clinicians (as in the case of medical internists constructing differential diagnoses [37]). However, while models such as Med-PaLM 2, Med-Gemini [50, 66], GPT-4 [45], and Health-Alpaca [31] excel at medical question answering and interpreting clinical data, their capabilities for interpreting personal health data is less well established.

Expanding LLMs to operate on modalities beyond just text has been a recent area of intense research, with prominent examples including but not limited to Flamingo [2], PaLI [13], GPT-4 [47], GPT-4v [48], Gemini 1.0 [21], and Gemini 1.5 [22]. The exploration of multimodal LLMs has also been extended to biomedical applications. Many models explore pairing one or multiple medical imaging modalities with language, including Med-Flamingo [40], LLaVA-Med [32], BiomedCLIP [68], MedBLIP [12], ELIXR [65], and others reviewed in further detail elsewhere [64]. Other models explore support for non-imaging medical modalities, including HeLM [7], Med-PaLM M [58], and Med-Gemini [50, 66]. While many of the earlier works focused primarily on medical question answering, there is increasing focus on report generation and other long-form responses.

Evaluation of long-form text is challenging [33] but is critical to ensure practical utility of LLMs in realistic settings. Similar to our efforts to generate case studies of personal health coaching scenarios, MedAlign introduced a dataset for evaluating LLMs on relevant clinical tasks [17] and demonstrated frequent misalignment between question answering performance and realistic task performance.

**Discriminative and Generative Personal Health** Wearable sensors can help people realize meaningful changes in their health, such as helping to increase the amount of physical activity they engage in [16]. Moreover, when done thoughtfully and in an evidence-based manner, it is generally accepted that helping individuals derive insights from their data could increase the frequency of engaging in beneficial health behaviors. In the field of mobile health research [25, 56], traditional methodologies have predominantly centered around specialized, predictive models for defined classification tasks, such as predicting heart rate [49], energy expenditure [19], blood pressure [6], and other vital signs, or classifying diseases using machine learning models tailored to specific predictive purposes such as atrial fibrillation detection [43] and improving objective rehabilitation monitoring [9]. More recently, LLMs have been shown be an effective base model to ground physiological and behavior time-series data and make meaningful inferences with zero-shot inference and few-shot learning across a wide variety of personal health tasks [31, 35]. In general, these methods use textual representations of sensor data to inform health metrics or predict health states. In contrast, our work with PH-LLM expands model utility from only predicting health states to also providing coherent, contextual, and potentially prescriptive outputs that depend on complex health behaviors. While traditional models operate within the confines of specific, often binary or multinomial, outcome prediction, PH-LLM interprets and generates recommendations based on health behaviors, providing a more interactive and interpretive utility. This evolution from predictive modeling to generative reasoning set out our contribution in bridging quantitative data interpretation with qualitative, contextually-rich output, facilitating a better experience of digital health interaction and personal health data utilization.

## 5 Conclusion

We developed an LLM fine-tuned from Gemini (PH-LLM) to perform a variety of tasks relevant to setting and achieving individual personal health goals. Our study shows that PH-LLM is capable of integrating passively-acquired objective data from wearable devices into personalized insights, potential causes for observed behaviors, and recommendations to improve sleep hygiene and fitness outcomes. After fine-tuning from the highly capable Gemini Ultra 1.0, which already displays aggregate performance approaching that of experts in fitness, PH-LLM demonstrated significantly improved use of domain knowledge and personalization of relevant user information for sleep

insights. Consistent with its strong performance on long-form case studies, we showed that PH-LLM accurately answers technical sleep and fitness multiple choice questions, and contextualize the benchmark performance of PH-LLM in these datasets with performance of multiple experts in the same tasks. Finally, we demonstrated the ability of PH-LLM to use a multimodal encoder that natively integrates time-series health behavior data as input tokens to predict subjective outcomes in sleep with performance on par with specialized models to predict the same outcomes.

**Acknowledgements**

We thank the Fitbit research community participants for making this research possible. We thank the sleep and fitness experts who developed case study responses and evaluated candidate model responses for their dedication, effort, and detailed feedback on multiple model iterations. Contributing sleep experts include Ben Graef, Timothy Wong, Thuan Dang, Suzanne Gorovoy, Narayan Krishnamurthy, and Michelle Jonelis. Contributing fitness experts include Jarod Spraggins, Allison Hetrick, Jonas Hannon, Max Knight, Nolan Dozier, Laura Grissom, and Justin Leach. We thank Hulya Emir-Farinas, Farhad Hormozdiari, and Joëlle Barral for feedback and discussions that significantly improved the work. We also thank Sami Lachgar, Lauren Winer, Maggie Shiels, Lee Gardner, Noa Tal, Annisah Um'rani, Oba Adewunmi, and Archit Mathur for their valuable insights, technical support, and feedback during our research.

**Competing Interests**

This study was funded by Google LLC. All authors are employees of Alphabet and may own stock as part of the standard compensation package.

## References


[1] S. Abbaspourazad, O. Elachqar, A. C. Miller, S. Emrani, U. Nallasamy, and I. Shapiro. Large-scale training of foundation models for wearable biosignals. *arXiv preprint arXiv:2312.05409*, 2023.

[2] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

[3] T. Althoff, R. Sosič, J. L. Hicks, A. C. King, S. L. Delp, and J. Leskovec. Large-scale physical activity data reveal worldwide activity inequality. *Nature*, 547(7663):336–339, 2017.

[4] American Board of Internal Medicine (ABIM). Sleep Medicine Blueprint Certification Examination (CERT), 2024. Accessed on August 27, 2024.

[5] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, et al. PaLM 2 Technical Report. *arXiv preprint arXiv:2305.10403*, 2023.

[6] T. Arakawa. Recent research and developing trends of wearable sensors for detecting blood pressure. *Sensors*, 18(9):2772, 2018.

[7] A. Belyaeva, J. Cosentino, F. Hormozdiari, K. Eswaran, S. Shetty, G. Corrado, A. Carroll, C. Y. McLean, and N. A. Furlotte. Multimodal llms for health grounded in individual-specific data. In *Workshop on Machine Learning for Multimodal Healthcare Data*, pages 86–102. Springer, 2023.

[8] BoardVitals. BoardVitals: Medical board review and continuing medical education, 2024. Accessed on August 27, 2024.

[9] P. Bonato. Advances in wearable technology and applications in physical medicine and rehabilitation. *Journal of NeuroEngineering and Rehabilitation*, 2(2):1–4, 2005.

[10] D. J. Buysse. Sleep Health: Can We Define It? Does It Matter? *Sleep*, 37(1):9–17, 2014.

[11] J.-P. Chaput, C. Dutil, R. Featherstone, R. Ross, L. Giangregorio, T. J. Saunders, I. Janssen, V. J. Poitras, M. E. Kho, A. Ross-White, et al. Sleep timing, sleep consistency, and health in adults: a systematic review. *Applied Physiology, Nutrition, and Metabolism*, 45(10):S232–S247, 2020.

[12] Q. Chen, X. Hu, Z. Wang, and Y. Hong. MedBLIP: Bootstrapping language-image pre-training from 3D medical images and texts. *arXiv preprint arXiv:2305.10799*, 2023.

[13] X. Chen, X. Wang, S. Changpinyo, A. Piergiovanni, P. Padlewski, D. Salz, S. Goodman, A. Grycner, B. Mustafa, L. Beyer, et al. PaLI: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.

[14] C.-H. Chiang and H.-y. Lee. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, 2023.

[15] G. T. Doran. There's a smart way to write managements's goals and objectives. *Management Review*, 70(11), 1981.

[16] T. Ferguson, T. Olds, R. Curtis, H. Blake, A. J. Crozier, K. Dankiw, D. Dumuid, D. Kasai, E. O'Connor, R. Virgara, et al. Effectiveness of wearable activity trackers to increase physical activity and improve health: a systematic review of systematic reviews and meta-analyses. *The Lancet Digital Health*, 4(8):e615–e626, 2022.

[17] S. L. Fleming, A. Lozano, W. J. Haberkorn, J. A. Jindal, E. Reis, R. Thapa, L. Blankemeier, J. Z. Genkins, E. Steinberg, A. Nayak, et al. MedAlign: a clinician-generated dataset for instruction following with electronic medical records. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22021–22030, 2024.

[18] M. Fogelholm. Physical activity, fitness and fatness: relations to mortality, morbidity and disease risk factors. A systematic review. *Obesity Reviews*, 11(3):202–221, 2010.

[19] D. Fuller, E. Colwell, J. Low, K. Orychock, M. A. Tobin, B. Simango, R. Buote, D. Van Heerden, H. Luan, K. Cullen, et al. Reliability and validity of commercially available wearable devices for measuring steps, energy expenditure, and heart rate: systematic review. *JMIR mHealth and uHealth*, 8(9):e18694, 2020.

[20] I. R. Galatzer-Levy, D. McDuff, V. Natarajan, A. Karthikesalingam, and M. Malgaroli. The capability of large language models to measure psychiatric functioning. *arXiv preprint arXiv:2308.01834*, 2023.

[21] Gemini Team, Google. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[22] Gemini Team, Google. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[23] W. J. Gordon, A. Landman, H. Zhang, and D. W. Bates. Beyond validation: getting health apps into clinical practice. *npj Digital Medicine*, 3(1):14, 2020.

[24] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

[25] E. Guillodo, C. Lemey, M. Simonnet, M. Walter, E. Baca-García, V. Masetti, S. Moga, M. Larsen, H. Network, J. Ropars, et al. Clinical applications of mobile health wearable–based sleep monitoring: systematic review. *JMIR mHealth and uHealth*, 8(4):e10733, 2020.

[26] K. L. Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48, 2008.

[27] K. L. Gwet. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014.

[28] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[29] C.-K. Kao and D. M. Liebovitz. Consumer mobile health apps: current state, barriers, and future directions. *PM&R*, 9(5):S106–S115, 2017.

[30] D. M. Katz, M. J. Bommarito, S. Gao, and P. Arredondo. GPT-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254, 2024.

[31] Y. Kim, X. Xu, D. McDuff, C. Breazeal, and H. W. Park. Health-LLM: Large language models for health prediction via wearable sensor data. *arXiv preprint arXiv:2401.06866*, 2024.

[32] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024.

[33] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[34] I. W. Lin, A. Sharma, C. M. Rytting, A. S. Miner, J. Suh, and T. Althoff. IMBUE: improving interpersonal effectiveness through simulation and just-in-time feedback with human-language model interaction. *arXiv preprint arXiv:2402.12556*, 2024.

[35] X. Liu, D. McDuff, G. Kovacs, I. Galatzer-Levy, J. Sunshine, J. Zhan, M.-Z. Poh, S. Liao, P. Di Achille, and S. Patel. Large language models are few-shot health learners. *arXiv preprint arXiv:2305.15525*, 2023.

[36] D. McDuff, A. Barakat, A. Winbush, A. Jiang, F. Cordeiro, R. Crowley, L. E. Kahn, J. Hernandez, N. B. Allen, et al. The Google Health Digital Well-Being Study: Protocol for a Digital Device Use and Well-Being Study. *JMIR Research Protocols*, 13(1):e49189, 2024.

[37] D. McDuff, M. Schaekermann, T. Tu, A. Palepu, A. Wang, J. Garrison, K. Singhal, Y. Sharma, S. Azizi, K. Kulkarni, et al. Towards accurate differential diagnosis with large language models. *arXiv preprint arXiv:2312.00164*, 2023.

[38] M. A. Merrill, A. Paruchuri, N. Rezaei, G. Kovacs, J. Perez, Y. Liu, E. Schenck, N. Hammerquist, J. Sunshine, S. Tailor, K. Ayush, H.-W. Su, Q. He, C. Y. McLean, M. Malhotra, S. Patel, J. Zhan, T. Althoff, D. McDuff, and X. Liu. Transforming wearable data into health insights using large language model agents. *arXiv preprint arXiv:2405.12345*, 2024.

[39] J. G. Meyer, R. J. Urbanowicz, P. C. Martin, K. O'Connor, R. Li, P.-C. Peng, T. J. Bright, N. Tatonetti, K. J. Won, G. Gonzalez-Hernandez, et al. Chatgpt and large language models in academia: opportunities and challenges. *BioData Mining*, 16(1):20, 2023.

[40] M. Moor, Q. Huang, S. Wu, M. Yasunaga, Y. Dalmia, J. Leskovec, C. Zakka, E. P. Reis, and P. Rajpurkar. Med-Flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023.

[41] National Institutes of Health (NIH). Patient-Reported Outcomes Measurement Information System (PROMIS), 2024. Accessed on August 27, 2024.

[42] National Strength and Conditioning Association (NSCA). Certified Strength and Conditioning Specialist (CSCS) exam, 2024. Accessed on August 27, 2024.

[43] S. Nemati, M. M. Ghassemi, V. Ambai, N. Isakadze, O. Levantsevych, A. Shah, and G. D. Clifford. Monitoring and detecting atrial fibrillation using wearable technology. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3394–3397. IEEE, 2016.

[44] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz. Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

[45] H. Nori, Y. T. Lee, S. Zhang, D. Carignan, R. Edgar, N. Fusi, N. King, J. Larson, Y. Li, W. Liu, et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.

[46] B. Öhlin, P. Nilsson, J.-Å. Nilsson, and G. Berglund. Chronic psychosocial stress predicts long-term cardiovascular morbidity and mortality in middle-aged men. *European Heart Journal*, 25(10):867–873, 2004.

[47] OpenAI. Gpt-4 technical report, 2024.

[48] OpenAI (2023). GPT-4V(ision) technical work and authors. `https://cdn.openai.com/contributions/gpt-4v.pdf`, 2023. Accessed: 2024-05-13.

[49] J. Parak and I. Korhonen. Evaluation of wearable consumer heart rate monitors based on photopletysmography. In *2014 36th annual international conference of the IEEE engineering in medicine and biology society*, pages 3670–3673. IEEE, 2014.

[50] K. Saab, T. Tu, W.-H. Weng, R. Tanno, D. Stutz, E. Wulczyn, F. Zhang, T. Strother, C. Park, E. Vedadi, et al. Capabilities of Gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

[51] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al. BLOOM: A 176B-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

[52] A. Sharma, K. Rushton, I. W. Lin, T. Nguyen, and T. Althoff. Facilitating self-guided mental health interventions through human-language model interaction: A case study of cognitive restructuring. *arXiv preprint arXiv:2310.15461*, 2023.

[53] A. Sharma, K. Rushton, I. W. Lin, D. Wadden, K. G. Lucas, A. Miner, T. Nguyen, and T. Althoff. Cognitive reframing of negative thoughts through human-language model interaction. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

[54] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[55] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

[56] S. R. Steinhubl, E. D. Muse, and E. J. Topol. The emerging field of mobile health. *Science Translational Medicine*, 7(283):283rv3–283rv3, 2015.

[57] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[58] T. Tu, S. Azizi, D. Driess, M. Schaekermann, M. Amin, P.-C. Chang, A. Carroll, C. Lau, R. Tanno, I. Ktena, et al. Towards generalist biomedical AI. *arXiv preprint arXiv:2307.14334*, 2023.

[59] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[60] S. G. Wannamethee, A. G. Shaper, and M. Walker. Changes in physical activity, mortality, and incidence of coronary heart disease in older men. *The Lancet*, 351(9116):1603–1608, 1998.

[61] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le. Finetuned language models are zero-shot learners. *International Conference on Learning Representations*, 2022.

[62] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[63] M. Wornow, Y. Xu, R. Thapa, B. Patel, E. Steinberg, S. Fleming, M. A. Pfeffer, J. Fries, and N. H. Shah. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1):135, 2023.

[64] H. Xiao, F. Zhou, X. Liu, T. Liu, Z. Li, X. Liu, and X. Huang. A comprehensive survey of large language models and multimodal large language models in medicine. *arXiv preprint arXiv:2405.08603*, 2024.

[65] S. Xu, L. Yang, C. Kelly, M. Sieniek, T. Kohlberger, M. Ma, W.-H. Weng, A. Kiraly, S. Kazemzadeh, Z. Melamed, et al. ELIXR: Towards a general purpose x-ray artificial intelligence system through alignment of large language models and radiology vision encoders. *arXiv preprint arXiv:2308.01317*, 2023.

[66] L. Yang, S. Xu, A. Sellergren, T. Kohlberger, Y. Zhou, I. Ktena, A. Kiraly, F. Ahmed, F. Hormozdiari, T. Jaroensri, et al. Advancing multimodal medical capabilities of Gemini. *arXiv preprint arXiv:2405.03162*, 2024.

[67] L. Yu, D. J. Buysse, A. Germain, D. E. Moul, A. Stover, N. E. Dodds, K. L. Johnston, and P. A. Pilkonis. Development of short forms from the PROMIS™ sleep disturbance and sleep-related impairment item banks. *Behavioral Sleep Medicine*, 10(1):6–24, 2012.

[68] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

# Appendix

## Table of Contents

# A Broader Impact

A primary overarching goal for developing models specific to personal health is to be able to improve long-term health outcomes through effective behavior change and maintenance of healthy habits. Neither of these tasks is explicitly evaluated here, and remain important areas for future work. While the performance of PH-LLM on the tasks presented here is encouraging, we caution that much work remains to be done to ensure LLMs are reliable, safe, and equitable in personal health applications. Further reducing confabulations, considering an individual's unique health circumstances not captured by sensor information alone, and ensuring alignment of the training data with real-world distributions are a subset of important research areas that warrant further attention.

# B Limitations

Our work has several limitations. First, the distribution of case study rubric ratings were skewed quite high, making differentiation across models and expert responses challenging. While some case study sections and evaluation rubric principles did show significant differentiation, further training of expert raters to increase inter-rater reliability or adjudicating existing responses could increase signal strength of model performance. Second, owing to inter-rater variability, we chose to have each expert rate all responses for a given case study. While this made direct comparison of candidate responses straightforward, it introduced the potential for experts to identify expert vs model responses based on style or other non-material factors, and thus introduce conscious or unconscious biases into ratings. Third, we observed that despite improvements in referencing and integrating user data into insights, confabulations or incorrect referencing of user data still occasionally occurred. Addressing and preventing these issues will be critical to ensure the safe and effective deployment of these technologies into user-facing features. Promising progress is being made through active research on agentic workflows that critique and correct candidate responses [38]. Fourth, the case studies were sampled broadly across demographics (sleep) or to identify common patterns in active individuals (fitness), but may not be a representative sample of the population nor exhaustively explore the sleep and fitness concerns affecting individuals. Fifth, our exploration of multimodal encoding of sensor data explored a small fraction of the design space owing to the relatively small dataset with paired outcome data and our purposeful restriction to samples with nearly complete sensor data. Further exploration of self-supervised pre-training on raw waveforms and granularly aggregated sensor features may yield richer representations of individuals that can be effectively purposed toward personal health outcome predictions [1] that expand beyond just sleep metrics and address challenges arising from a sparse and heterogeneous mix of available sensor features. We anticipate that future large datasets with paired outcome data will enable non-linear interactions across features to be learned effectively to improve predictive power.

Despite the above limitations, we have demonstrated here that the Gemini family of models are imbued with substantial health knowledge, and we can effectively fine-tune Gemini Ultra 1.0 to improve performance across multiple outcomes relevant for personal health. The results from this study represent an important step toward LLMs that deliver personalized information and recommendations that support individuals to achieve their health goals.

# C Methods

## C.1 Base model selection

In order to start from the most capable base model, we performed automated evaluation of several Gemini candidate model sizes and a medical LLM on the professional exam questions. The candidate models were Gemini Nano 1.0, Gemini Pro 1.0, Gemini Ultra 1.0 [21], and MedPaLM-2. Gemini Ultra 1.0 consistently produced the best accuracy on professional examinations (Figures E.1 and E.2).

## C.2 Base model prompting on case studies

Since Gemini Ultra 1.0 was the most accurate model on professional examinations, suggesting it has appropriate domain knowledge in the areas of sleep and fitness, we explored the performance of this model on case studies. We prompted Gemini Ultra 1.0 by summarizing guidelines given to the experts for dataset creation. For example, the sleep experts generally were asked to follow the RU-SATED format (Routine, Sleep Quality, Alertness, Timing, Efficiency, and Duration) [10] to generate sleep

insights. In order to give Gemini Ultra 1.0 the best shot at answering case studies, we similarly prompt it to follow the RU-SATED format and provide an explanation of what metrics should be used to assess each dimension (see Table D.1-D.10 for details). We note that each case study consisted of multiple sections representing different queries and responses: three sections for sleep case studies (insights, etiology, recommendations) and five sections for fitness case studies (demographics, training load, sleep, health metrics, and the assessment). Since each section represented a different aspect of the case study, we developed prompts specifically for each section. Tables D.1-D.5 show the prompts for sleep case studies and Tables D.6-D.10 show the prompts for fitness case studies. For sections that synthesized results from previous sections, i.e., the etiology and recommendation sections in sleep case studies, and the assessment section in fitness case studies, we substituted the model answers from previous sections into the prompt (see Table D.10 for an example).

### C.3 Training PH-LLM on case studies

We fine-tuned Gemini Ultra 1.0 on the dataset of coaching recommendations and call this model PH-LLM. We use the case studies from the training, validation, and test sets for model training and selection (457 case studies for sleep and 300 case studies for fitness). For each of the sleep and fitness domains, we randomly split the dataset into separate training, validation, and test splits using a 70:15:15 ratio. We used the same prompts that were given to the baseline model to form prompt-response pairs for model tuning. Since each section was treated as a separate example, this resulted in 1,371 prompt-response pairs for sleep and 1,500 prompt-response pairs for fitness across the training, validation, and test sets (Figure D.1A,B).

Typically, LLMs are trained on mixture of tasks [61]. Here we fine-tuned the model on a 1:1 mixture of sleep and fitness prompt-response pairs. Within the fitness prompt-response pairs, we chose to upsample higher quality case studies by a 2:1 ratio, where higher quality case studies were defined as those that underwent additional rounds of quality control by the fitness experts.

The model was fine-tuned for a maximum of 1500 steps with a global batch size of 4 using linear warm-up over 50 steps and cosine decay. We used a learning rate of $2.5 \times 10^{-7}$, weight decay of $1 \times 10^{-2}$, and a learning rate decay minimum ratio of 0.1. We saved model checkpoints every 50 steps. For our final model candidate, we chose the first checkpoint after the model had been trained for at least one epoch (this checkpoint also had a relatively low log perplexity).

### C.4 Training PH-LLM for patient-reported outcomes

To train PH-LLM to predict PROs from wearable data, we followed the methodology developed in HeLM [7]. Wearable data for each user was stored as a matrix in which the rows represent wearable measurement devices and the columns represent measurements at a specific time. In our case, we had 20 device measurements measured once over 15 days for each sample in the dataset. Next, we encoded this data by computing the mean and variance across days, and z-scoring the results using the training data as a reference. This yielded a new "encoded" matrix of $20 \times 2$ where columns correspond to a measure's mean and variance. The encoded data matrix was projected into the token embedding space of PH-LLM via a multilayer perceptron (MLP) adapter with three hidden layers (sizes 1,024, 4,096, and 1,024) and an output of 2 tokens. The resulting set of tokens were provided to PH-LLM as a prefix to the text input, which included a text representation of all input fields in their native form (e.g., steps per day; not z-scored). We prompted the model to predict a specific binary outcome (e.g., "I am satisfied with my sleep - 'yes' or 'no"'). An example of the corresponding text prompt is shown in Table F.20. The adapter was trained via backpropagation while keeping PH-LLM weights frozen.

We compared these adapter-based predictions to text-only predictions using both zero-shot and few-shot prompting. For zero-shot, the prompt format was identical to the adapter-based prediction except the adapter token prefix was omitted. For few-shot, as many complete examples as could fit within the context window (up to seven) were included as exemplars. For all three models, the positive and negative outcomes were scored by computing the log likelihood for each outcome.

Text prompts that included using only mean results, and both mean and variance, were explored (while always including both mean and variance in the input to the MLP adapter). Since performance was not appreciably different (data not shown), we omitted the variance encoding to enable more in-context examples to be passed as textual context to PH-LLM.

As a separate comparison, we fitted logistic regression models separately for each binary outcome, in which the predictors were the same mean and variance computed across 15 days of sensor data. For both the MLP adapter and logistic models we trained using the shared training set, selected the best model according to ROAUC in the validation set and then present results using the final holdout set.

### C.5 Expert grading of case study responses

While evaluation against MCQs and PROs can be performed by comparing model predictions to gold-standard structured responses and numerical values, respectively, the case studies involve longer-form outputs.

In order to evaluate these longer-form case study responses, the domain experts (including all individuals involved in creating the case study responses) were asked to evaluate three responses written to each case study: one by Gemini Ultra 1.0, one by PH-LLM, and one by a domain expert. Each domain expert was assigned evaluations randomly to case studies for which they did not write the expert response. The domain experts evaluated each case study response based on a custom rubric that quantifies incorporation of user data, appropriate personalization based on user data, use of expert domain knowledge, evidence of confabulations or unwarranted assumptions, potential for harm, readability, and overall quality. The complete set of evaluation questions for the case studies is provided in Appendix D.2.

Evaluation cases were fully distributed across the primary group of experts based on availability during the research project's evaluation period. A portion of the evaluation case studies were additionally assigned to the rest of the available domain experts to ensure on-schedule, thorough completion of the evaluation dataset.

Both the creation of expert written case study responses and the evaluation of all 3 types of responses were performed on an internal health data labeling platform that adheres to data privacy and security best practices and design principles. It handles labeling task creation, scheduling and assignment, answer storage as well as front-end visualization and labeling through its web application. It supports highly customizable viewers for multiple data modalities including medical images and text reports. We customized the HTML viewer to display long-form case studies comprising figures, tables, and text, in an effective and intuitive manner.

### C.6 Automatic evaluation of case study responses

Though expert grading of case study responses was our primary mechanism for assessing model performance, it is a time-consuming process that scales poorly. This makes it challenging to iterate on model improvements since sending all checkpoints to human raters is prohibitively expensive. Automated evaluation (AutoEval) allows us to obtain a quick–though potentially less accurate than human evaluation–signal that can be used during model development by using secondary models to perform this rating task [14]. In this section, we describe our approach for curating a case study response rating dataset, fine-tuning AutoEval models capable of rating candidate models, and using AutoEval to select promising models that are then sent to expert raters for human feedback.

While exploring different modeling mechanisms, we performed an initial round of expert grading using the rubrics and procedure described in Appendix C.5 for 50 expert-generated case studies from each vertical across three response sources: experts, an untuned Gemini Ultra 1.0 model, and a fine-tuned Gemini Pro 1.0 model. We then split these studies into vertical-specific training and validation splits containing roughly 80% ($N = 38$) and 20% ($N = 12$) of case studies, respectively. Splits were structured such that samples rated by a given expert were evenly distributed between sets. All ratings associated with a given case study were included in that split, resulting in $N = 6,552$ total ratings across case study sections and evaluation principles for sleep ($N = 4,872$ train; $N = 1,596$ validation) and $N = 9,331$ for fitness ($N = 7,138$ train; $N = 2,193$ validation). Using these ratings and the corresponding case study data and responses, we constructed LLM prompts and targets matching the format described in Table D.20 (see Table D.21 for a full example). Prompts included a description of the rating task objective for the given case study section, a summary of data describing the case study, the principle being assessed, and the principle's Likert scale options. Each target was the expert-generated rating followed by the rating's Likert option text description (e.g., for a "No incorrect domain knowledge" principle rating of 5, the target is "5. No incorrect domain knowledge references exist.").

We fine-tuned Gemini Pro 1.0 models using LoRA [28] across a variety of vertical-specific data mixtures, including all ratings for a vertical and all ratings from a single rater. All AutoEval modeling experiments used a fixed set of hyperparameters, varying only the training data mixture: a LoRA rank of 4 on attention heads, a constant learning rate of $2 \times 10^{-5}$, a global batch size of 32, and a maximum of 20 epochs for the given training mixture. We present results for the following data mixtures:

1. All ratings in either the fitness or sleep verticals ("All").
2. All ratings from the lowest variance rater in the fitness ("Fitness Primary B") or sleep ("Sleep Primary D") verticals, where variance is calculated across all ratings from that expert.
3. All ratings from the highest variance rater in the fitness ("Fitness Primary C") or sleep ("Sleep Primary C") verticals.

An untuned Gemini Pro 1.0 model served as a baseline. We generated model predictions by scoring the likelihood of each Likert option given the input prompt, converted these scores into five-class multinomial probabilities, and chose the option with the largest probability score. We selected candidate AutoEval models using a combination of log perplexity loss and Spearman's rank correlation between predictions and the ground truth ratings in the validation dataset.

Given case study responses from candidate PH-LLM models trained using the procedure described in Appendix C.3, we used the same scoring procedure above to automatically rate model outputs across case study sections and evaluation principles. We used these ratings in conjunction with non-expert feedback to filter candidate models for full human expert evaluation. We then used the resulting ratings to further evaluate the performance of our final AutoEval models.

### C.7 Statistical analyses

Confidence intervals (95%) were determined via bootstrapping with 1,000 iterations. Statistical significance of expert ratings was determined using a two-sided Wilcoxon rank-sum test with false-discovery rate (Benjamini-Hochberg) correction when multiple sections or multiple evaluation principles were analyzed together. All p-values refer to p-values after FDR correction. For each p-value, we report the test statistic $Z$ and the effect size $r = Z/\sqrt{N}$, where $N$ is the total sample size of the test.

# D Coaching Recommendations Case Studies

## D.1 Case study dataset creation

### D.1.1 Additional details on creation of sleep case studies

Our study utilized de-identified data from individuals who provided consent for use of their data for research purposes. For sleep case studies, in order to ensure a representative sample across different demographics (age and gender), we considered 64 different demographic groups, determined by a combination of 32 different age buckets (13-20 years old, 20-80 years old with each group within this range spanning two years, 80 years old and above) and 2 gender buckets (male, female).

**Daily Sleep Metrics:** The daily sleep metrics contain up to 29 days of daily sleep metrics. The metrics are: date, day of the week, sleep score (0-100), light sleep (hh:mm), REM sleep (hh:mm), deep sleep (hh:mm), sleep duration (hh:mm), fall asleep time, wake time after sleep onset (hh:mm), sleep efficiency, fraction of sleep goal, number of times the individual woke up, heart rate (bpm), nap duration (min), number of naps, and wake up time. See Table D.4 for an example.

**Aggregated Daily Sleep Statistics:** Generally, these statistics included an aggregated metric (e.g. average, median, standard deviation, count) over all the days, the percentile that the aggregated metric is in as compared to other individuals within the same demographic group, minimum value over all the days, maximum value over all the days, as well as 5th and 95th percentiles of the aggregated metric as compared to other individuals within the same demographic group. In some instances, such as bedtime, the metrics were computed separately for all days, weekdays only, and weekends only to understand weekday versus weekend patterns. See Table D.5 for an example.

The experts composed responses across the following sections:

**Insights:** Implicitly this section was aimed at answering the question of "What are some sleep-related insights based on my data?" The sleep medicine expert examined the data and provided an interpretation of whether a data point might represent an atypical sleep pattern. The experts were asked to systematically review each case to provide a holistic assessment of the user's sleep patterns. To do so, Fitbit sleep metrics were assessed according to the validated RU-SATED framework (Routine, Sleep Quality, Alertness, Timing, Efficiency, and Duration) to generate sleep insights [10].

**Etiology:** Implicitly this section answered the question of "What are the possible underlying causes that could explain the observed data?" The experts generally considered the contribution of circadian rhythm, homeostatic drive, psychophysiologic hyperarousal, and extrinsic factors and indicated their likelihood.

**Recommendations:** This section was generally designed to answer the question of "What can I do to improve my sleep?" The experts were asked to provide personalized recommendations to the individual that can help them improve their sleep by addressing potential causes identified in the etiology section. The experts were instructed to utilize best practices in goal-setting using the SMART framework.

### D.1.2 Additional details on creation of fitness case studies

For fitness case studies, the individuals from de-identified cohort who provided consent for use of their data for research purposes, were sampled. In order to ensure the fitness case studies contain sufficient activity for interesting training readiness analysis, we sampled individuals who had data for at least 16 days with minimum mean active zone minutes of 45 minutes and with at least 2 logged exercises. In addition, we considered periods of days that contained noticeable changes in heart rate variability, resting heart rate, respiratory rate, sleep, and periods with runs. The experts considered the following sections and data in their analysis:

**Demographics:** (age, gender, height, weight, body mass index). The experts considered the demographics data and commented on whether any precautions should be taken when recommending a fitness program.

**Training Load:** The experts were provided with a detailed table capturing daily metrics over the past 30 days, including day of the week, date, minutes spent in fat-burn, cardio, and peak zones, training impulse (TRIMP), and number of steps (Table D.11). Additionally, we provided aggregated statistical analyses such as means, ranges, acute TRIMP (7-day total training load), chronic TRIMP (28-day average acute training load), Acute-Chronic Workload Ratio (ACWR), and metrics specific

to each exercise entry (Table D.12). For ease of analysis, in addition to the table, daily TRIMP values were visualized in a barplot (Figure 2D).

**Sleep Metrics:** The experts assessed the individual's sleep as it relates to fitness recovery. A table of daily sleep metrics such as bedtime, wake time, sleep time, awake time, deep sleep, REM sleep, and sleep score was given to the experts for analysis (Table D.13). For ease of analysis, some of the daily sleep metrics were also visualized as a graph (Figure 2E) They were also given aggregated metrics including means, standard deviations, and z-scores indicating the difference in metrics between the most recent 3 days and the past 28 days to identify recent trends.

**Health Metrics:** A table and a graph of daily resting heart rate, heart rate variability (HRV), and respiratory rate over the past 30 days was given to the experts to assess recovery and stress (Figure 2F, Table D.15). The experts were also given aggregate metrics such as means, standard deviations, ranges, and z-scores indicating the difference in metrics between the most recent day and the past 28 days (Table D.17).

To simulate feedback from the user about their subjective state, the experts were also given synthetically (LLM) generated subjective readiness to workout (e.g. "feeling fatigued") and muscle soreness (e.g. "manageable soreness"). For examples, see Tables D.18 and D.19.

**Assessment & Recommendation:** The information from the previous sections was used to provide a summary of the most important insights. These insights along with synthetically generated user input on subjective readiness and muscle soreness (e.g., Tables D.18 and D.19) were used to inform an assessment of how ready the individual is to perform a workout today on the scale of 1 to 5. The experts also provided fitness recommendations to the individual (Figure 2G).

### D.1.3 Holistic View of Case Study Creation.

Each set of domain experts consisted of "primary" and "secondary" contributors to case study response creation and evaluation. This categorization was based on an expert's general availability to contribute to the research project on a weekly basis throughout its duration; "primary" contributors had more involvement and higher volumes of case study response creation and evaluation than "secondary" contributors. The grouping was primarily used for research project operations planning and scheduling. The level of domain expertise was similar across the two groups. Each vertical also included a clinical lead with extensive background in sleep medicine for the sleep vertical and sport and exercise medicine for the fitness vertical. The clinical lead oversaw case study development and provided feedback and quality control to the set of domain experts.

To generate the dataset used for training, validation, and testing, we first prompted the Gemini family of models with the data for each section in order to generate baseline model (Gemini Ultra 1.0) responses (Figure D.1A). The experts then reviewed the responses and rewrote them as needed. The dataset also underwent multiple rounds of quality control engaging the experts and clinical leads. Separately, to generate the holdout dataset, the experts wrote the responses from scratch (without any LLM assistance). This was done to ensure a more clear comparison between experts and the model during evaluation.

In total, we created 350 case studies for fitness (300 case studies for the training, validation, and test set and 50 case studies for the holdout set) and 507 case studies for sleep (457 case studies for the training, validation, and test set and 50 case studies for the holdout set).

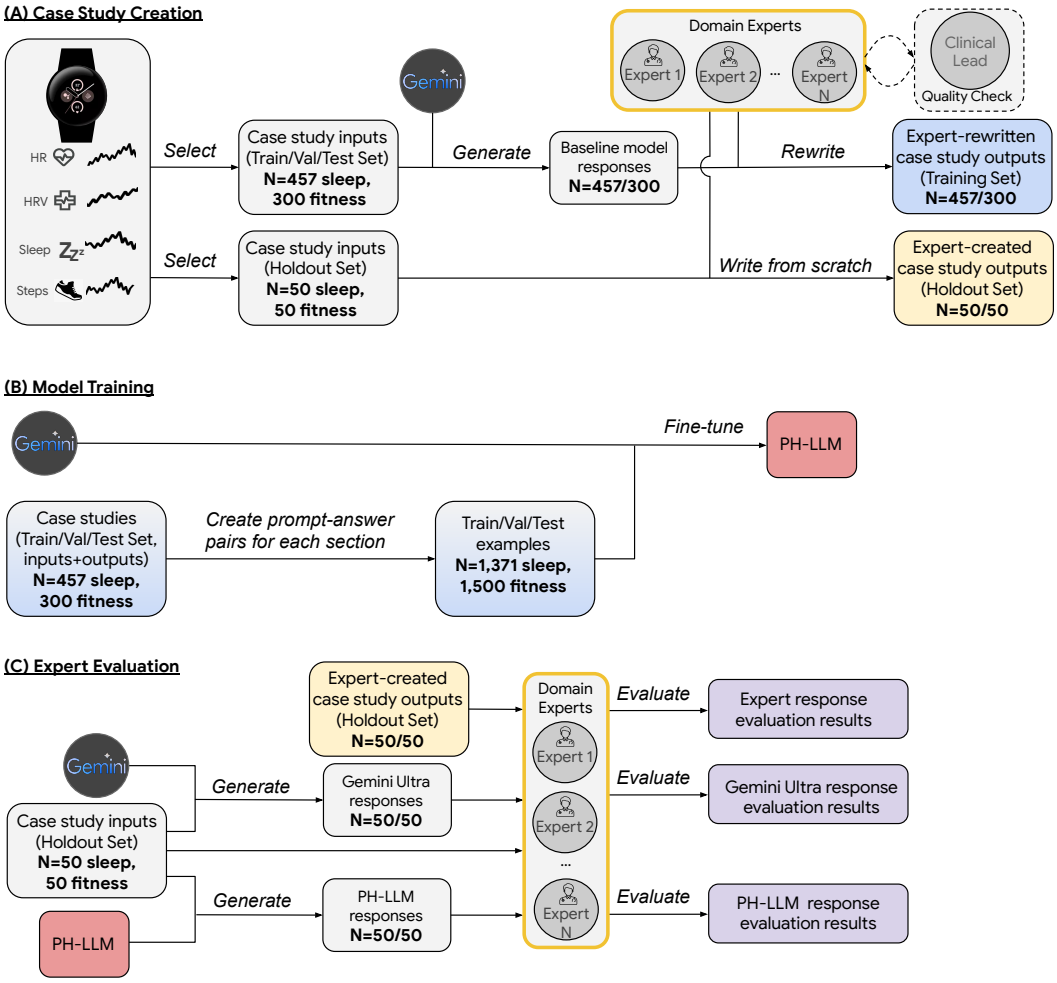



Figure D.1: **Case Study Creation, Curation, and Evaluation Workflow.** Case studies were selected from a large set of anonymized, consented production data. **(A)** Two sets of case studies were generated. To facilitate rapid development of high-quality answers, the train/validation/test set of case studies had candidate responses generated by Gemini, which were then edited and rewritten by domain experts. To enable comparison of human and model-derived responses, the holdout set had responses written solely by the domain experts. **(B)** For model training, each case study was split into multiple prompt/answer pairs based on how many sections the case study had (N=3 for sleep with insights, etiology, and recommendations sections, N=5 for fitness with demographics, training load, sleep metrics, health metrics, and assessment sections, see Section C.3 for details) and Gemini Ultra 1.0 underwent full fine-tuning using those examples. **(C)** Expert evaluation was performed independently on the holdout dataset by the same set of domain experts responsible for generating the expert responses. For each case study in the holdout set, an expert who did not write the corresponding expert response graded all three candidate responses (expert-written response, Gemini Ultra 1.0 response, PH-LLM response).

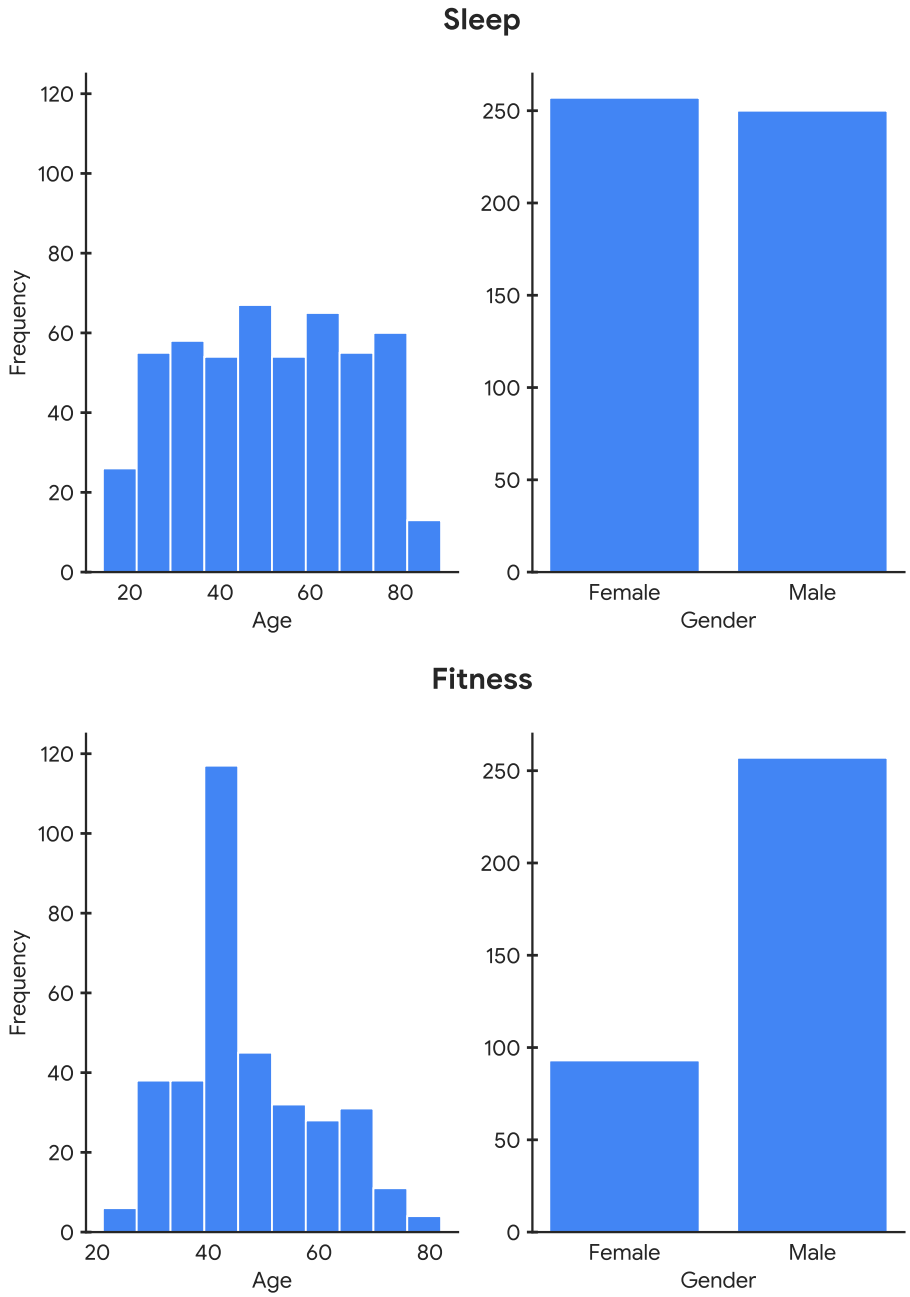



Figure D.2: **Distribution of case study ages and genders across the sleep and fitness verticals.**

Table D.1: **Prompt for sleep case studies insights section.**

Prompt for sleep case studies insights section

You are a sleep medicine expert. You are given the following sleep data.
The user is <gender>, <age> years old.
**Sleep logs:**
<sleep logs table (see Table D.4 for an example)>

**Sleep Summary**:
<sleep summary (see Table D.5 for an example)>

List the most important insights. Identify all of the patterns of data that are likely out of the preferred range. Make sure to consider various sleep health dimensions: Routine, Sleep Quality, Alertness, Timing, Efficiency, and Duration. Add a heading for each dimension. Optionally (only do this if extremely important) add a heading called Other for anything else that doesn't fit the above categories. For Routine, consider the average bedtime, wake time, midsleep point and standard deviations of these, focus on the consistency of the routine, not timing. For Sleep Quality, consider light sleep duration, deep sleep duration, REM sleep duration, sleep score, restlessness score, time to quality sleep, and wake time after sleep onset. For Alertness, consider the number of naps and nap length. For Timing, consider midsleep point, bedtime, wake time, make any comments on weekend vs. workday. For Efficiency, consider sleep efficiency, wake time after sleep onset, and time to quality sleep, describe how they compare to similar users. For Duration, consider average sleep duration, weekend vs. workday sleep durations and standard deviations, describe how they compare to similar users. When determining whether a metric is normal or abnormal, always provide the corresponding percentile. Avoid generic statements. Avoid incorrect knowledge, inconsistencies and contradictions. Don't mention "the user". Talk like you're speaking directly to someone. Be concise.
# Sleep insights report

Table D.2: **Prompt for sleep case studies etiology section.**

Prompt for sleep case studies etiology section

You are a sleep medicine expert. You are given the following sleep data.
The user is <gender>, <age> years old.
**Sleep Summary**:
<sleep summary (see Table D.5 for an example)>

Based on the data, we can get the following insights:
<insights response>
What are the underlying causes? Make sure to consider the following causes: Circadian rhythm, Homeostatic drive, Psychophysiologic hyperarousal, and Extrinsic factors. Order the causes from most to least relevant. Identify the likelihood of the causes (e.g. unlikely, possible, very likely). Cite relevant data and insights, for example, "consistently low sleep efficiency despite normal sleep durations suggests low homeostatic drive". Avoid diagnosing health conditions. Avoid providing recommendations. Avoid generic statements. Avoid incorrect knowledge, inconsistencies and contradictions. Don't mention "the user". Talk like you're speaking directly to someone. Be concise.
# Causes report

Table D.3: **Prompt for sleep case studies recommendations section.**

Prompt for sleep case studies recommendations section

You are a sleep medicine expert. You are given the following sleep data.
The user is <gender>, <age> years old.
**Sleep Summary**:
<sleep summary (see Table D.5 for an example)>

Based on the data, we can get the following insights:
<insights response>
Causes:
<etiology response>
What recommendation(s) can you provide to help this user improve their sleep? Tie recommendations to the very likely and possible causes, for example, "Recommendations to address Circadian rhythm". Tie recommendations to user's sleep data such as average bedtime, average wake time, and number of naps, and recommend a goal bedtime and wake time based on their data. The recommendations should be time-bound, for example for the next week or the next month. Write one short question to ask the user in order to better understand their sleep. Avoid assumptions regarding the trainee's lifestyle or behavioral choices. Avoid generic statements. Avoid incorrect knowledge, inconsistencies and contradictions. Don't mention "the user". Talk like you're speaking directly to someone. Be concise.
# Recommendations report

Table D.4: **Abridged example of sleep logs table for a particular individual used in sleep case studies.** For brevity, only seven days are shown.

Abridged example of sleep logs table for a particular individual used in sleep case studies

| Date | Day of Week | Sleep Score | Light Sleep (hh:mm) | REM Sleep (hh:mm) | Deep Sleep (hh:mm) | Sleep Duration (hh:mm) | Fall Asleep Time | Wake after Sleep Onset (hh:mm) | Efficiency | Fraction of Sleep Goal | Wakeup Count | Heart Rate (bpm) | Nap Duration (min) | Naps | Wake Time |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| <year-month-day> | Thursday | 71.0 | 04:24 | 00:59 | 00:40 | 06:04 | 00:05 | 00:04 | 0.88 | 0.76 | 4.0 | 58.0 | 0.0 | 0.0 | 06:13 |
| <year-month-day> | Friday | 72.0 | 03:13 | 01:07 | 01:03 | 05:24 | 00:38 | 00:08 | 0.85 | 0.68 | 8.0 | 58.0 | 88.0 | 1.0 | 06:10 |
| <year-month-day> | Saturday | 87.0 | 05:08 | 01:51 | 02:00 | 09:00 | 03:02 | 00:10 | 0.87 | 1.12 | 9.0 | 58.0 | 0.0 | 0.0 | 12:12 |
| <year-month-day> | Sunday | 83.0 | 05:16 | 01:49 | 01:41 | 08:47 | 03:54 | 00:15 | 0.86 | 1.10 | 15.0 | 58.0 | 0.0 | 0.0 | 12:56 |
| <year-month-day> | Monday | 68.0 | 04:21 | 00:50 | 00:42 | 05:54 | 00:07 | 00:08 | 0.85 | 0.74 | 8.0 | 58.0 | 0.0 | 0.0 | 06:09 |
| <year-month-day> | Monday | 64.0 | 01:29 | 00:27 | 00:51 | 02:48 | 16:10 | 00:05 | 0.85 | 0.35 | 5.0 | 58.0 | 0.0 | 0.0 | 19:03 |
| <year-month-day> | Tuesday | 70.0 | 01:18 | 00:43 | 00:50 | 02:52 | 03:42 | 00:02 | 0.87 | 0.36 | 2.0 | 59.0 | 0.0 | 0.0 | 06:36 |
| <year-month-day> | Wednesday | 72.0 | 03:19 | 01:14 | 01:02 | 05:36 | 00:17 | 00:09 | 0.83 | 0.70 | 9.0 | 58.0 | 0.0 | 0.0 | 06:02 |
| <year-month-day> | Wednesday | 71.0 | 01:41 | 00:43 | 00:35 | 03:00 | 16:22 | 00:00 | 0.86 | 0.38 | 0.0 | 58.0 | 0.0 | 0.0 | 19:22 |

Table D.5: **Abridged example of sleep summary for a particular individual used in sleep case studies.** Stratified features report overall statistics as well as stratified by workday vs weekend, and include bedtime, wake time, midsleep point, sleep duration, and sleep score. Unstratified features include time to quality sleep, wake time after sleep onset, sleep efficiency, light sleep duration, deep sleep duration, REM sleep duration, and restlessness score. Nap length and total number of naps are also reported.

**Abridged example of sleep summary for a particular individual used in sleep case studies.**

Average bedtime is 00:26
Average bedtime is in the 65th percentile
Earliest bedtime is 16:10
Latest bedtime is 06:22
Bottom 5th percentile of similar users' average bedtimes is 21:25
Top 95th percentile of similar users' average bedtimes is 03:07

Bedtime standard deviation is 03:34
Bedtime standard deviation is in the 94th percentile

Average bedtime on the weekend is 01:35
Average bedtime on the weekend is in the 72nd percentile
Earliest bedtime on the weekend is 16:58
Latest bedtime on the weekend is 06:22
Bottom 5th percentile of similar users' average bedtimes on the weekend is 21:45
Top 95th percentile of similar users' average bedtimes on the weekend is 03:28

Bedtime standard deviation on the weekend is 03:46
Bedtime standard deviation on the weekend is in the 92nd percentile

Average bedtime on a workday is 23:58
Average bedtime on a workday is in the 60th percentile
Earliest bedtime on a workday is 16:10
Latest bedtime on a workday is 06:22
Bottom 5th percentile of similar users' average bedtimes on a workday is 21:10
Top 95th percentile of similar users' average bedtimes on a workday is 03:14

Bedtime standard deviation on a workday is 03:28
Bedtime standard deviation on a workday is in the 94th percentile

Median bedtime on a workday is 00:38
Median bedtime on a workday is in the 73rd percentile
Bottom 5th percentile of similar users' median bedtimes on a workday is 21:04
Top 95th percentile of similar users' median bedtimes on a workday is 03:23

...
Average time to quality sleep is 00:33
Average time to quality sleep is in the 92nd percentile
Shortest time to quality sleep is 00:04
Longest time to quality sleep is 01:23
Bottom 5th percentile of similar users' average times to quality sleep is 00:13
Top 95th percentile of similar users' average times to quality sleep is 00:35

...
Average nap length is 129
Average nap length is in the 92nd percentile

Total number of naps is 4

Table D.6: **Prompt for fitness case studies demographics section.**

**Prompt for fitness studies demographics section**

You are a NSCA and ACSM board-certified fitness trainer who specializes in athlete training performance and recovery.
Age: <age>
Height: <height>
Weight: <weight>
BMI: <BMI>
Gender: <gender>

Are there any special precautions that should be taken into account when recommending a fitness program to avoid injury? Comment if the trainee has exceptional demographics (e.g. very old, very high BMI, very low BMI) that require special considerations. Write a single sentence. Avoid mentioning diseases.

Table D.7: **Prompt for fitness case studies training load section.**

Prompt for fitness case studies training load section

The following section shows some of the trainee's recent activity metrics including the active zone minutes: Fat burn zone (50% heart rate reserve), Cardio zone (70% heart rate reserve), and Peak zone (85% heart rate reserve.)
**Daily activity metrics:**

<table of daily activity metrics (see Table D.11 for an example)>
Today is <day of the week> <year-month-day>.

**Here are some aggregate statistics for the last 30 days:**
<aggregate statistics of daily activity metrics (see Table D.12 for an example)>

Analyze the trainee's recent activity metrics, aggregate statistics for the last 30 days, and most recent exercise logs. Assess the following: Training Load Trends, Intensity, Duration, Frequency, Rest Periods, Acute-Chronic-Workload Ratio (ACWR), Recent Activity Levels, and Significant Workouts. For Training Load Trends, consider mean moderate activity per day, mean vigorous activity per day, comment on balance between moderate and vigorous activity. For Intensity, consider the most recent exercise logs, assess time in fat-burn zone (moderate intensity), time in cardio zone (vigorous intensity), time in peak zone (peak intensity), and state whether the workouts overall reached each zone, consider the daily activity metrics and assess the TRIMP values. For Duration, consider the most recent exercise logs and list the lowest and highest duration as a range. For Frequency, consider the most recent exercise logs, and check on which days of the week there is a workout. For Rest Periods, consider the daily activity metrics table and see if some days have very low to zero TRIMP - these are also rest periods, comment on the number of rest days and which days of the week. For Acute-Chronic-Workload Ratio, consider acute TRIMP, chronic TRIMP, see if acute TRIMP is higher than chronic TRIMP and state what it means in terms of training load, consider Acute-Chronic Workload Ratio (ACWR) and state what it means for recovery. ACWR values above 1.5 reflect a significant increase in training load and may result in a higher risk of injury. ACWR values of less than 0.7 indicate that the trainee has had a significant decrease in training load and may be at risk of detraining. For Recent Activity Levels and Significant Workouts, consider the most recent exercise logs and note any recent significant workouts that are related to changes in the training load metrics, consider the daily activity metrics and highlight days with highest TRIMP and explain their importance.

Note: Remember to avoid readiness assessments, avoid recommendations, avoid making up data, and stay directly aligned with the provided data.
- Base all observations and insights on the provided data.
- Avoid generic advice.
- Refrain from making up data or giving general advice not rooted in the data.
- Avoid assumptions regarding the trainee's lifestyle or behavioral choices.
- Do not elaborate on anything not contained within the data tables.
- Do not compute or reference complex mathematical calculations like correlation coefficients.
- When explaining the numerical difference, refrain from inventing any calculations if you are not certain about them.
- Use markdown to structure the response.
- Use an observation/insight format:
* **Observation:** A factual observation from the data.
* **Insight:** The implication of the observation in the context of the user's health.
- Group the observation/insights into appropriate sections.
# Training load report

Table D.8: **Prompt for fitness case studies sleep section.**

Prompt for fitness case studies sleep section

These are the trainee's **recent sleep metrics:**

<table of sleep metrics for fitness case studies (see Table D.13 for an example)>
Today is <day of the week> <year-month-day>.

**Here are some aggregate statistics for the last 30 days:**
<aggregate statistics of sleep metrics (see Table D.14 for an example)>

- Assess the following aspects of trainee's sleep based on metrics:
* Sleep Schedule: bedtimes and wake-times
* Sleep Duration: sleep duration metrics
* Sleep Quality: sleep score. Excellent sleep score is 90 to 100. Good sleep score is 80 to 89. Fair sleep score is 60 to 79. Poor sleep score is less than 60.
* Today's Sleep: Comment on today's values and compare them to the aggregate statistics for the last 30 days. Make this comment only if sleep duration Z-score or sleep score Z-score is less than -2, comment that this indicates significantly worse recent sleep in the last 3 days compared to the monthly average sleep duration and low final readiness assessment is recommended . Make this comment only if sleep duration Z-score or sleep score Z-score is more than 2, comment that this indicates significantly improved recent sleep in the last 3 days compared to the monthly average sleep duration.

- Base all observations and insights on the provided data.
- Avoid generic advice.
- Refrain from making up data or giving general advice not rooted in the data.
- Avoid assumptions regarding the trainee's lifestyle or behavioral choices.
- Do not elaborate on anything not contained within the data tables.
- Do not compute or reference complex mathematical calculations like correlation coefficients.
- When explaining the numerical difference, refrain from inventing any calculations if you are not certain about them.
- Be very concise.
- Avoid ## Recommendations.
- Avoid ## Overall Insights
- Use markdown to structure the response.
- Use an observation/insight format:
* **Observation:** A factual observation from the data.
* **Insight:** The implication of the observation in the context of user's health.
- Group the observation/insights into appropriate sections.
# Sleep report

Table D.9: **Prompt for fitness case studies health metrics section.**

Prompt for fitness case studies health metrics section

Here are some of the trainee's **daily health metrics for the past month:**

<table of health metrics over past 30 days (see Table D.15 for an example)>

Here are some of the trainee's **daily health metrics for the past week:**

<table of health metrics over past week (see Table D.16 for an example)>
Today is <day of the week> <year-month-day>.
**Here are some aggregate statistics for the last 30 days:**
<aggregate statistics of health metrics (see Table D.17 for an example)>

- Examine patterns for each health metric:
* Resting heart rate
* Heart rate variability
* Respiratory rate
- For each metric:
* Comment on the general baseline values.
* Comment on any trends/changes or consistency/typical/normal range of the metrics in the latest week compared to the month.
* Comment on today's values and compare them to the baseline and recent trends.
* Place emphasis on recent values in relation to long-term aggregated data.
- The Z-scores are number of standard deviations today's values are from the trainee's monthly baseline. Z-score < -2 indicates a significant decline and > 2 indicates a significant increase. Do not refer to the Z-scores directly.

Note: The goal is to extract as much actionable information as possible from the metrics, particularly in the context of understanding someone's recovery state.- Base all observations and insights on the provided data.
- Avoid generic advice.
- Refrain from making up data or giving general advice not rooted in the data.
- Avoid assumptions regarding the trainee's lifestyle or behavioral choices.
- Do not elaborate on anything not contained within the data tables.
- Do not compute or reference complex mathematical calculations like correlation coefficients.
- When explaining the numerical difference, refrain from inventing any calculations if you are not certain about them.
- Be concise.
- Avoid ## Overall insights.
- Use markdown to structure the response.
- Use an observation/insight format:
* **Observation:** A factual observation from the data.
* **Insight:** The implication of the observation in the context of user's health.
- For example use the following template:
## Resting Heart Rate
**Observation:**
**Insight:**
## Heart rate variability
**Observation:**
**Insight:**
## Respiratory rate
**Observation:**
**Insight:**

# Health report

Table D.10: **Prompt for fitness case studies readiness assessment section.**

Prompt for fitness case studies readiness assessment section

Use the following observations and insights to personalize the response below.
<demographics response>

<training load response>

<sleep metrics response>

<health metrics response>

The trainee has also provided the following qualitative feedback:
<subjective readiness>
<muscle soreness>

Based on the above observations and insights, determine the trainee's readiness to workout today. Use the following template and provide 1-2 bullet points for each section:
**Load**
**Sleep**
**Health Metrics**
**Subjective Readiness + Muscle Soreness**
**Readiness Score**
* X/5
* Explanation:
**Fitness Recommendations for Today**
**Followup Question**

For Load, Sleep, Health Metrics, and Subjective Readiness + Muscle Soreness, provide a short summary of the most important observations and insights, referencing any data, that are relevant to trainee's readiness to train today. Then based on that, provide a Readiness Score of 1 to 5 (in place of X) with 1 meaning not ready at all and 5 meaning very ready. 3 means the trainee may be ready with adaptation to their workout. Provide an explanation for why this score was chosen. Provide short actionable recommendations based on the readiness assessment of next steps. Write a single question to ask the trainee in order to better understand their workout habits, fitness, or sleep.
# Readiness summary report

Table D.11: **Abridged example of daily activity metrics table for a particular individual used in fitness case studies.** For brevity, only seven days of activity are shown.

Abridged example of daily activity metrics table for a particular individual used in fitness case studies

| Day of the week | date | Fat-burn zone minutes | Cardio zone minutes | Peak zone minutes | TRIMP | Steps |
|---|---|---|---|---|---|---|
| Wednesday | <year-month-day> | 15.0 | 27.0 | 0.0 | 62.0 | 16200 |
| Thursday | <year-month-day> | 19.0 | 23.0 | 1.0 | 62.0 | 9900 |
| Friday | <year-month-day> | 6.0 | 0.0 | 0.0 | 6.0 | 5950 |
| Saturday | <year-month-day> | 20.0 | 0.0 | 0.0 | 20.0 | 11210 |
| Sunday | <year-month-day> | 1.0 | 0.0 | 0.0 | 1.0 | 8160 |
| Monday | <year-month-day> | 7.0 | 0.0 | 0.0 | 7.0 | 13120 |
| Tuesday | <year-month-day> | 12.0 | 0.0 | 0.0 | 12.0 | 15490 |

Table D.12: **Abridged example of aggregated daily activity metrics table for a particular individual used in fitness case studies.** Full exercise logs contain at most 10 most recent exercise logs. Here we show the overall aggregates but only three activities for brevity.

Abridged example of aggregated daily activity metrics table for a particular individual used in fitness case studies.

Mean moderate activity per day (Fat-burn): 12.3 mins
Mean vigorous activity per day (Cardio and Peak): 12.7 mins
TRIMP ranges from 0 to 124
Acute TRIMP (7-day total training load): 346
Chronic TRIMP (28-day average acute training load): 235
Acute-Chronic Workload Ratio (ACWR): 1.5

These are exercise logs from most recent days.
Walk on Wednesday <year-month-day>
Duration: 17 mins
Average Heart Rate: 98 bpm
Time in Fat-burn zone: 18 mins
Time in Cardio zone: 0 mins
Time in Peak zone: 0 mins
Distance: 0 km
TRIMP that day: 47.0

Walk on Wednesday <year-month-day>
Duration: 11 mins
Average Heart Rate: 88 bpm
Time in Fat-burn zone: 8 mins
Time in Cardio zone: 0 mins
Time in Peak zone: 0 mins
Distance: 0 km
TRIMP that day: 47.0

Treadmill on Thursday <year-month-day>
Duration: 46 mins
Average Heart Rate: 140 bpm
Time in Fat-burn zone: 7 mins
Time in Cardio zone: 13 mins
Time in Peak zone: 14 mins
Distance: 5 km
TRIMP that day: 53.0
Average workout duration: 19.2 mins
Workout duration ranges from 10 to 46 mins
Average heart rate ranges from 80 to 140 bpm

Table D.13: **Abridged example of sleep metrics table for a particular individual used in fitness case studies.** For brevity, only seven days are shown.

Abridged example of sleep metrics table for a particular individual used in fitness case studies.

| Day of the week | Date | Sleep start time | Sleep end (wake) time | Sleep time (hours) | Awake time (minutes) | Deep sleep (minutes) | REM sleep (minutes) | Sleep score |
|---|---|---|---|---|---|---|---|---|
| Wednesday | <year-month-day> | 23:01 | 07:05 | 7 | 53 | 80 | 18 | 80 |
| Thursday | <year-month-day> | 22:48 | 07:17 | 7 | 49 | 94 | 17 | 84 |
| Friday | <year-month-day> | 22:43 | 07:12 | 7 | 77 | 61 | 13 | 71 |
| Saturday | <year-month-day> | 00:15 | 08:12 | 7 | 55 | 87 | 21 | 83 |
| Sunday | <year-month-day> | 01:11 | 09:33 | 7 | 62 | 86 | 15 | 74 |
| Monday | <year-month-day> | 23:16 | 07:31 | 7 | 57 | 104 | 19 | 86 |
| Tuesday | <year-month-day> | 22:13 | 04:04 | 4 | 55 | 41 | 13 | 64 |

Table D.14: **Example of aggregated sleep metrics table for a particular individual used in fitness case studies.**

Example of aggregated sleep metrics table for a particular individual used in fitness case studies.

Mean bedtime: 00:11
Mean wake-time: 07:35
Mean sleep duration: 6.3 hours
Standard deviation sleep duration: 1.3 hours
Sleep duration Z-score (recent days relative to month): -0.6
Mean sleep score: 76
Standard deviation sleep score: 9.1
Sleep score Z-score (recent days relative to month): -0.2

Table D.15: **Example of health metrics table over the past 30 days for a particular individual used in fitness case studies.**

Example of health metrics table for a particular individual used in fitness case studies.

| Day of the week | Date | Resting Heart Rate (bpm) | HRV RMSSD (ms) | Respiratory Rate (breaths/minute) |
|---|---|---|---|---|
| Wednesday | <year-month-day> | 53.0 | 27 | 14 |
| Thursday | <year-month-day> | 54.0 | 22 | 13 |
| Friday | <year-month-day> | 55.0 | 27 | 13 |
| Saturday | <year-month-day> | 56.0 | 23 | 15 |
| Sunday | <year-month-day> | 57.0 | 23 | 14 |
| Monday | <year-month-day> | 56.0 | 31 | 14 |
| Tuesday | <year-month-day> | 56.0 | 19 | 15 |
| Wednesday | <year-month-day> | 58.0 | NaN | NaN |
| Thursday | <year-month-day> | 61.0 | 17 | 15 |
| Friday | <year-month-day> | 64.0 | 13 | 15 |
| Saturday | <year-month-day> | 62.0 | 23 | 15 |
| Sunday | <year-month-day> | 63.0 | 16 | 15 |
| Monday | <year-month-day> | 62.0 | 26 | 14 |
| Tuesday | <year-month-day> | 60.0 | 28 | 14 |
| Wednesday | <year-month-day> | 61.0 | 17 | 15 |
| Thursday | <year-month-day> | 59.0 | 30 | 14 |
| Friday | <year-month-day> | 57.0 | 35 | 15 |
| Saturday | <year-month-day> | 58.0 | 25 | 16 |
| Sunday | <year-month-day> | 58.0 | 20 | 16 |
| Monday | <year-month-day> | 60.0 | 16 | 15 |
| Tuesday | <year-month-day> | 58.0 | 29 | 14 |
| Wednesday | <year-month-day> | 56.0 | 40 | 13 |
| Thursday | <year-month-day> | 54.0 | 41 | 14 |
| Friday | <year-month-day> | 56.0 | 28 | 15 |
| Saturday | <year-month-day> | 57.0 | NaN | NaN |
| Sunday | <year-month-day> | 60.0 | 17 | 16 |
| Monday | <year-month-day> | 62.0 | 15 | 15 |
| Tuesday | <year-month-day> | 65.0 | 19 | 16 |
| Wednesday | <year-month-day> | 67.0 | 16 | 16 |
| Thursday | <year-month-day> | 66.0 | 18 | 16 |

Table D.16: **Example of health metrics table for a particular individual used in fitness case studies.**

Example of health metrics table for a particular individual used in fitness case studies.

| Day of the week | Date | Resting Heart Rate (bpm) | HRV RMSSD (ms) | Respiratory Rate (breaths/minute) |
|---|---|---|---|---|
| Friday | <year-month-day> | 56.0 | 28 | 15 |
| Saturday | <year-month-day> | 57.0 | NaN | NaN |
| Sunday | <year-month-day> | 60.0 | 17 | 16 |
| Monday | <year-month-day> | 62.0 | 15 | 15 |
| Tuesday | <year-month-day> | 65.0 | 19 | 16 |
| Wednesday | <year-month-day> | 67.0 | 16 | 16 |
| Thursday | <year-month-day> | 66.0 | 18 | 16 |

Table D.17: **Example of aggregated health metrics table for a particular individual used in fitness case studies.**

| Example of aggregated health metrics table for a particular individual used in fitness case studies. |
|---|
| Mean Resting Heart Rate: 59 bpm<br>Standard deviation Resting Heart Rate: 3 bpm<br>Resting Heart Rate Z-score: 1.9<br>Mean HRV RMSSD: 24 ms<br>Standard deviation HRV RMSSD: 7 ms<br>HRV RMSSD Z-score: -0.8<br>Mean Respiratory Rate: 15 breaths/minute<br>Standard deviation Respiratory Rate: 0.83 breaths/minute<br>Respiratory Rate Z-score: 0.9<br>Past week:<br>Resting Heart Rate range: 56 to 67 bpm<br>HRV RMSSD range: 16 to 28 ms<br>Respiratory Rate range: 15 to 17 breaths/min |

Table D.18: **Example of synthetically-generated user input for subjective readiness to workout used in fitness case studies.**

| Example of synthetically-generated user input for subjective readiness to workout used in fitness case studies. |
|---|
| 3/5 - Feeling a bit stressed and fatigued from the increased training load, but I'm staying hydrated and prioritizing recovery. |

Table D.19: **Example of synthetically-generated user input for muscle soreness used in fitness case studies.**

| Example of synthetically-generated user input for muscle soreness used in fitness case studies. |
|---|
| Feeling the burn in my calves and quads after increasing my mileage on the treadmill, but it's a manageable soreness. |

### D.2 Case study evaluation rubrics

The case studies were graded using the rubric below. Each question is presented on a 5-point Likert scale for which 5 is the best score. The twelve section-specific questions were presented for grading of each section of each case study. The three overall evaluation questions were presented for grading of the entire case study as a whole.

**Section-specific evaluation questions**

**Q1. This section references all *important* user data needed.**

1. None of the important user data is referenced
2. There are some pieces of important user data referenced but most important user data is missing
3. About half of the important user data is referenced
4. Most of the important user data is referenced
5. All important user data is referenced

**Q2. This section does not reference *unimportant* user data.**

1. Only unimportant user data is referenced
2. Many unimportant user data references exist
3. Several unimportant user data references exist
4. A few unimportant user data references exist
5. No unimportant user data references exist

**Q3. This section does not reference *incorrect* user data, (e.g., hallucinated user data, incorrect variable, incorrect time period).**

1. Only incorrect user data is referenced
2. Many incorrect user data references exist
3. Several incorrect user data references exist
4. A few incorrect user data references exist
5. No incorrect user data references exist

**Q4. This section contains all *important* interpretations (aka personalization).**

1. None of the important interpretations are referenced
2. There are many important data interpretations missing
3. There are several important data interpretations missing
4. There are a few important data interpretations missing
5. All important data interpretations are present

**Q5. This section does not contain *unimportant* data interpretations (aka unimportant personalization).**

1. All of the data interpretations are unimportant
2. Many of the data interpretations are unimportant
3. Some of the data interpretations are unimportant
4. A few of the data interpretations are unimportant
5. None of the data interpretations are unimportant

**Q6. This section does not contain errors in its *important* interpretations, and correctly refuses to answer when such data is missing.**

1. All of the important data interpretations are incorrect

2. Many of the important data interpretations are incorrect
3. Some of the important data interpretations are incorrect
4. A few of the important data interpretations are incorrect
5. None of the important data interpretations are incorrect

**Q7. This section does not contain errors in its *unimportant* interpretations.**

1. All of the unimportant data interpretations are incorrect
2. Many of the unimportant data interpretations are incorrect
3. Some of the unimportant data interpretations are incorrect
4. A few of the unimportant data interpretations are incorrect
5. None of the unimportant data interpretations are incorrect

**Q8. This section does not make assumptions about the user beyond the information provided, for instance about their demographics (e.g race, ethnicity, health, lifestyle) or associated stereotypes.**

1. There are many assumptions present
2. There are several assumptions present
3. There are a few assumptions present
4. There is 1 assumption present
5. No assumptions are present

**Q9. This section contains evidence of *important* domain knowledge (e.g., mention of a relevant and/or correct fact for answering the question).**

1. No important domain knowledge is referenced
2. There are some pieces of important domain knowledge referenced but most data is missing
3. About half of the important domain knowledge is referenced
4. Most of the important user data is referenced
5. All important domain knowledge is referenced

**Q10. This section does not contain evidence of *unimportant* domain knowledge (e.g knowledge which has no use for the task objective).**

1. Only unimportant domain knowledge is referenced
2. Many unimportant domain knowledge references exist
3. Several unimportant domain knowledge references exist
4. A few unimportant domain knowledge references exist
5. No unimportant domain knowledge references exist

**Q11. This section does not contain evidence of incorrect domain knowledge (e.g factually incorrect or not accepted by expert consensus).**

1. Only incorrect domain knowledge is referenced
2. Many incorrect domain knowledge references exist
3. Several incorrect domain knowledge references exist
4. A few incorrect domain knowledge references exist
5. No incorrect domain knowledge references exist

**Q12. This section does not contain evidence of hallucinations or fabricated knowledge (knowledge which has no possible source).**

1. Only references to hallucinations or fabricated knowledge exists

2. Many references to hallucinations or fabricated knowledge exist
3. Several references to hallucinations or fabricated knowledge exist
4. A few references to hallucinations or fabricated knowledge exist
5. No references to hallucinations or fabricated knowledge exist

**Overall evaluation questions**

**Overall Q1. This case study does not contain information that could lead to harm.**

1. There is information that would likely cause the User serious harm
2. There is information that would likely cause the User some harm
3. There is information that may cause the User harm
4. There is potentially harmful information, but it is not significant
5. There is no information that could lead to harm

**Overall Q2. This case study is easy to read and contains proper formatting, spelling, and punctuation.**

1. The case is very difficult to read due to grammatical or formatting issues
2. There are many grammatical or formatting issues that affect readability
3. There are some grammatical or formatting issues that affect readability
4. There are a few grammatical or formatting issues, but is still easy to read
5. The case is easy to read and has no grammatical or formatting issues

**Overall Q3. What is the overall quality of this case study?**

1. Very Poor: the entire case needs to be rewritten
2. Poor: There are some highly significant errors present
3. Fair: The case could be improved
4. Good: Some slight improvements are possible
5. Excellent: No changes needed

### D.3 Additional case study performance evaluations

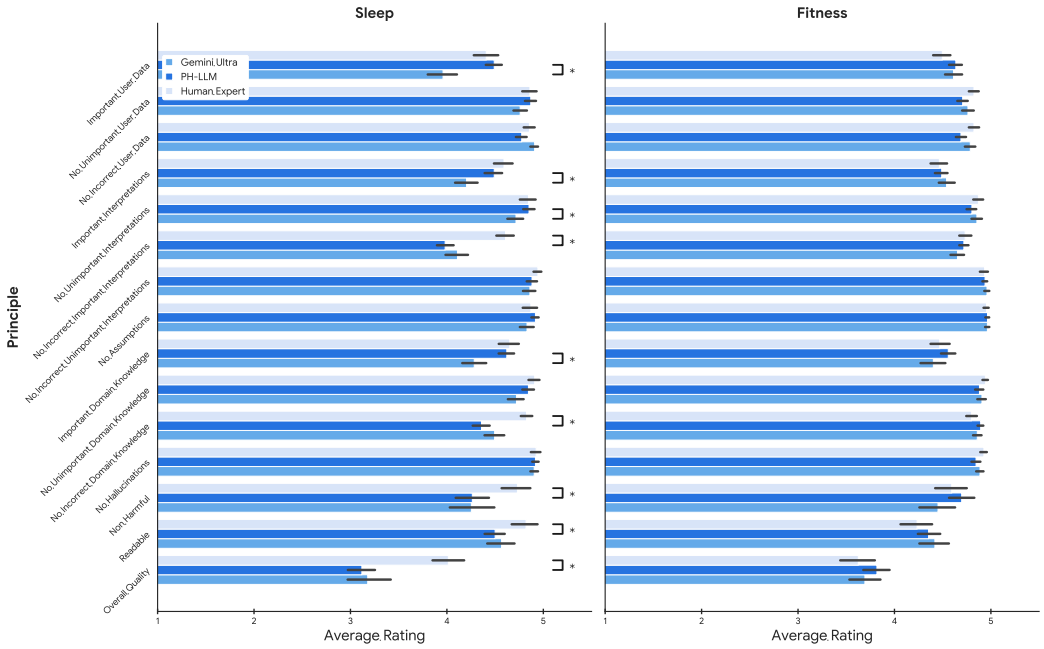



Figure D.3: **Sleep and fitness case study human evaluation results by principle.** Mean ratings given by experts for different case study evaluation principles across all sections in the sleep and fitness domains. The principles are ordered according to the rubric presented in Section D.2. "$*$" indicates a statistically significant difference between two response types after multiple hypothesis testing correction. Error bars represent 95% confidence intervals.

### D.4 Inter-rater expert agreement and rating speeds across primary and secondary raters

In order to assess agreement between raters and analyze differences between primary and secondary rater groups, we introduced a small amount of overlap across case study rating assignments within each vertical, resulting in anywhere from 78 to 1,428 paired ratings within a subset of raters. To evaluate inter-rater reliability, we employed several established metrics: raw counts of agreement, pairwise Spearman's rank correlation, Kendall's Coefficient of Concordance (Kendall's W), Weighted Cohen's Kappa, and Gwet's AC2. Spearman's correlation and Kendall's Coefficient of Concordance are metrics based on comparing ranks of expert ratings. These may be more conservative in the presence of many ties, as in our data with many ratings clustered around 4 and 5. Weighted Cohen's Kappa measures agreement between raters adjusted for chance agreement. The metric ranges from -1 to 1 with zero indicating that the agreement among raters is similar to chance. Weighted Cohen's Kappa can exhibit paradoxical behavior ("Kappa Paradox"), underestimating the true extent of agreement between raters [26] in imbalanced datasets such as ours. Gwet's AC1, and its weighted extension designed for ordinal data, Gwet's AC2, were designed to deal with class imbalance while adjusting for chance agreement [26, 27]. Based on both contingency tables showing raw agreement between raters (Figures D.4 and D.5) and measures of inter-rater reliability (Figures D.6 and D.7), we conclude moderate inter-rater reliability with conservative metrics like Weighted Cohen's Kappa being above 0, which indicates agreement beyond chance and with Gwet's AC2 ranging from 0.699 to 0.956.

We also generally observe that primary raters tend to have higher measures of agreement with one another than with secondary raters. However, due to low sample sizes, this difference is not significant (see Figures D.4 and D.5 for contingency tables and Figures D.6 and D.7 for agreement measures). We also measured the amount of time it took for each rater to rate all sections and principles for a given case study in minutes. We find that primary raters rate significantly faster than secondary raters in both verticals (Table D.23).

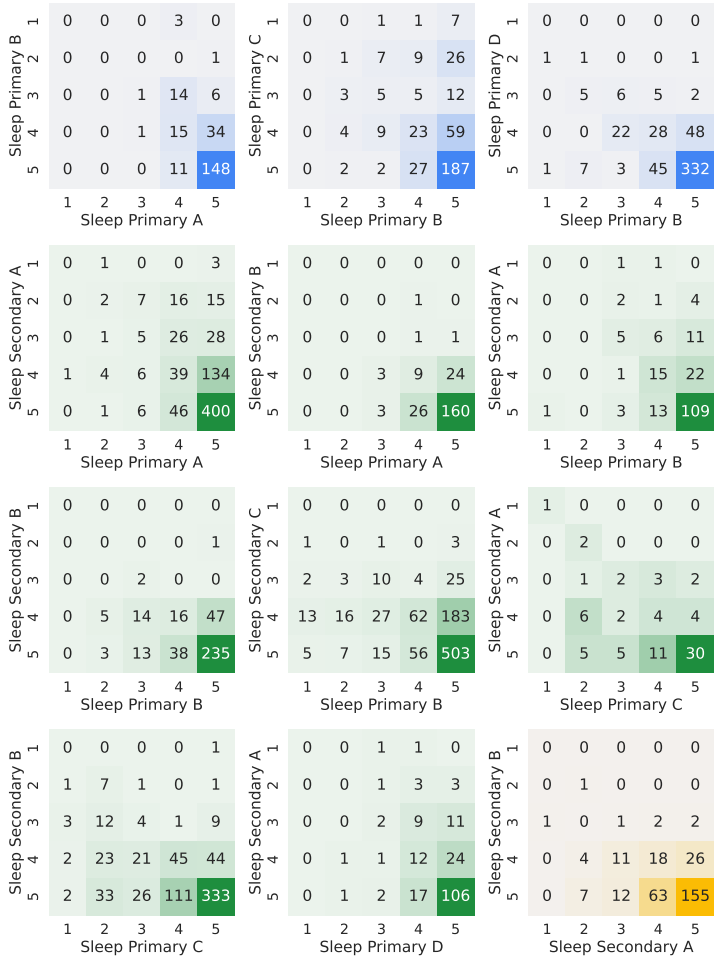



Figure D.4: **Contingency tables showing pairwise rating agreement between raters in the sleep vertical.** Counts are aggregated across all case studies, sections, and principles for each case study for which multiple ratings are available. Blue, primary vs primary raters. Green, primary vs secondary raters. Yellow, secondary vs secondary raters.

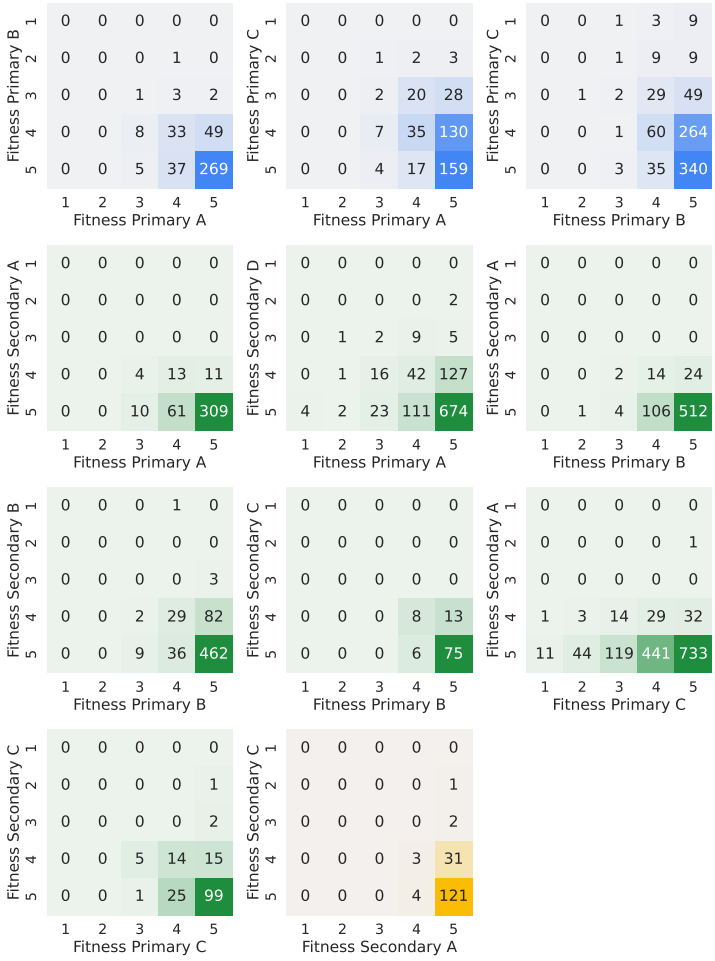



Figure D.5: **Contingency tables showing pairwise rating agreement between raters in the fitness vertical.** Counts are aggregated across all case studies, sections, and principles for each case study for which multiple ratings are available. Blue, primary vs primary raters. Green, primary vs secondary raters. Yellow, secondary vs secondary raters.

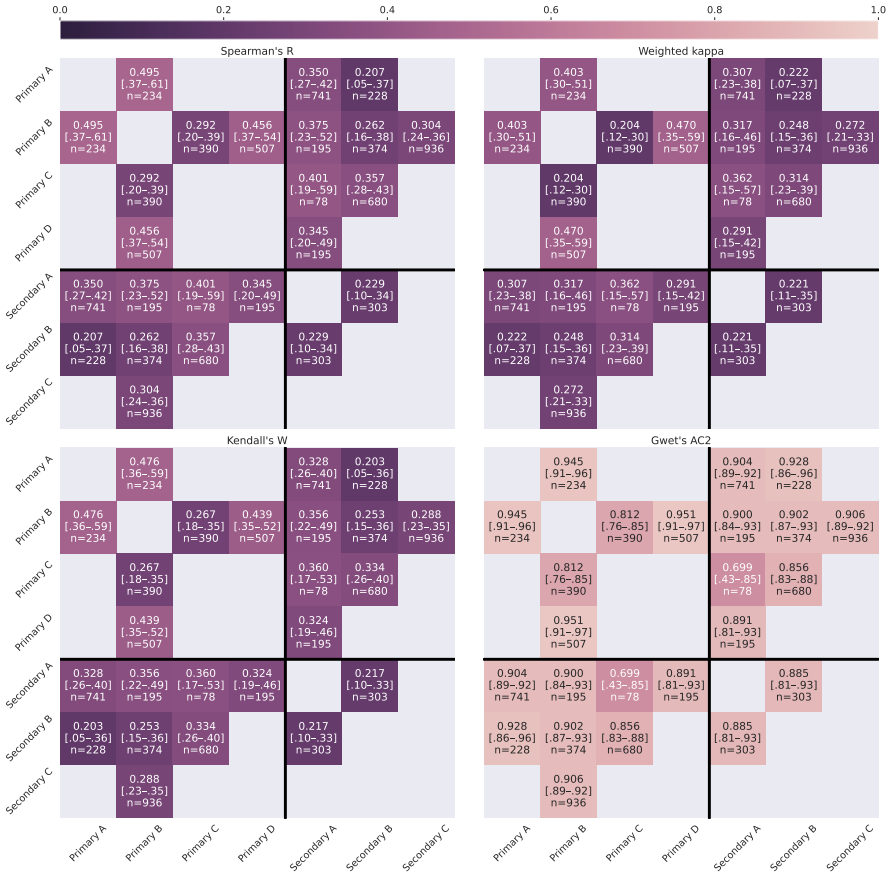



Figure D.6: **Pairwise Spearman's rank correlation, Weighted Cohen's Kappa, Kendall's Coefficient of Concordance (Kendall's W), and Gwet's AC2 measuring concordance between primary and secondary raters in the sleep vertical.** Metrics were computed using all ratings for each principle and section across case studies rated by more than one rater. The number of overlapping ratings is denoted by $n$. Mean metrics and 95% confidence intervals derived from 1,000 bootstrapping iterations are reported for each pair.

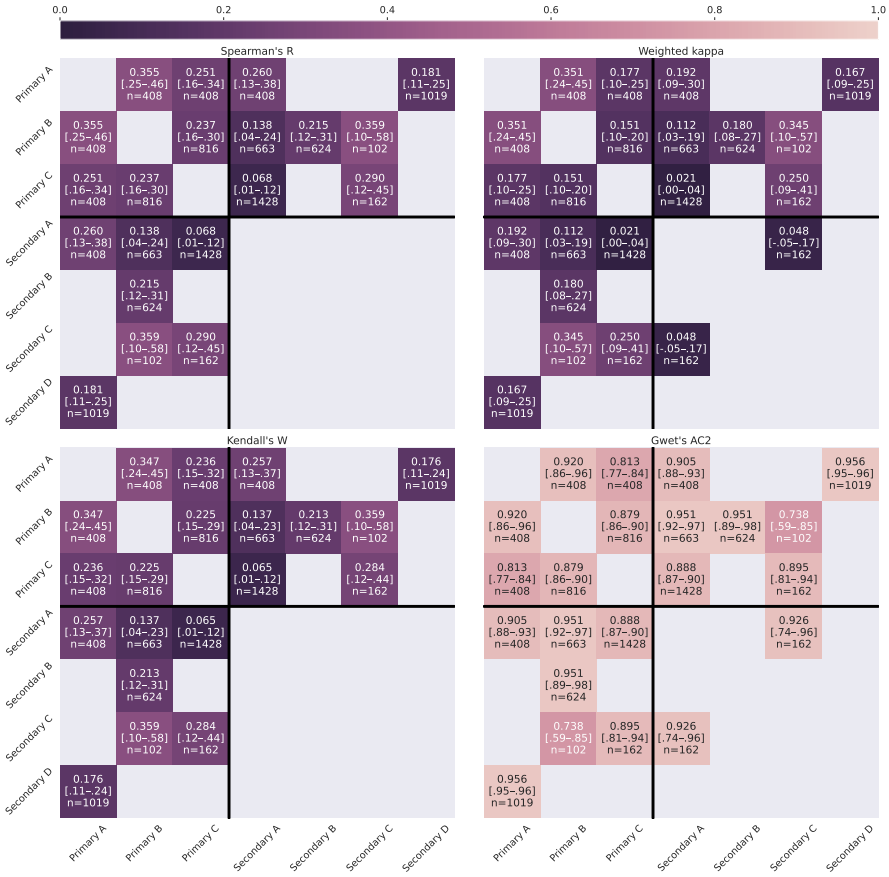



Figure D.7: **Pairwise Spearman's rank correlation, Weighted Cohen's Kappa, Kendall's Coefficient of Concordance (Kendall's W), and Gwet's AC2 measuring concordance between primary and secondary raters in the fitness vertical.** Metrics were computed using all ratings for each principle and section across case studies rated by more than one rater. The number of overlapping ratings is denoted by $n$. Mean metrics and 95% confidence intervals derived from 1,000 bootstrapping iterations are reported for each pair.

### D.5 Automatic evaluation of case studies

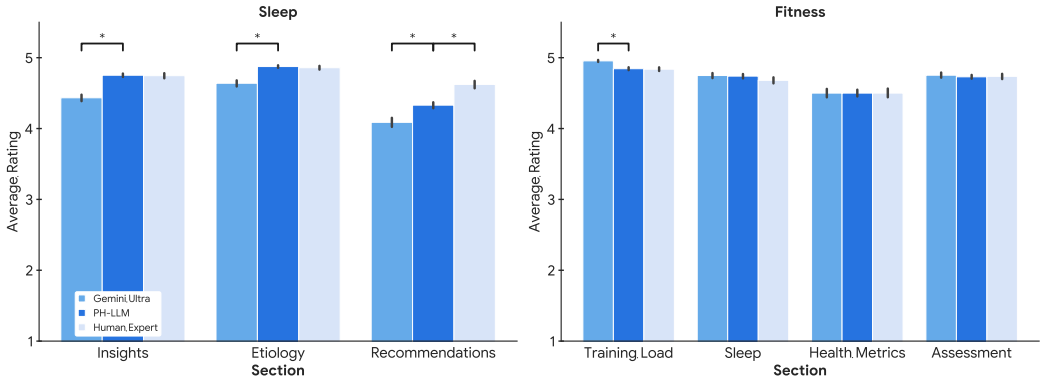



Figure D.8: **Case Study AutoEval Evaluation Results Using High-Variance Raters.** Mean ratings given by AutoEval models tuned using ratings from high variance raters for the case study subsections across the **(Left)** sleep and **(Right)** fitness domains. "$*$" indicates a statistically significant difference between two response types after multiple hypothesis testing correction. Error bars represent 95% confidence intervals.

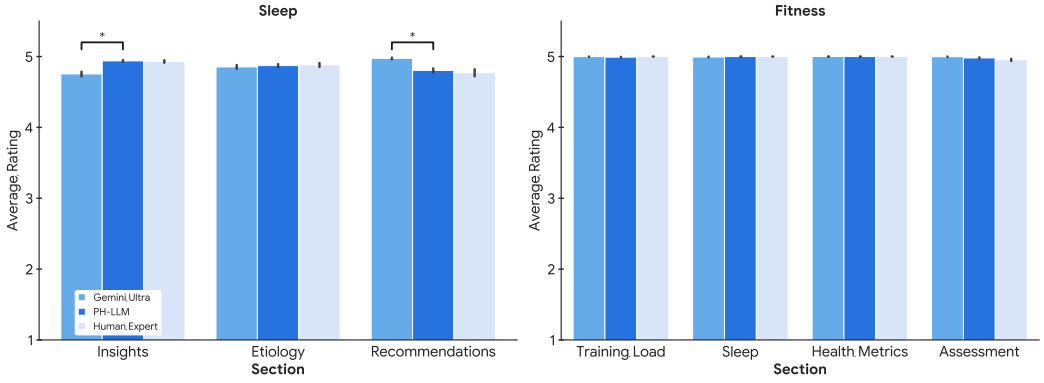



Figure D.9: **Case Study AutoEval Evaluation Results Using Low-Variance Raters.** Mean ratings given by AutoEval models tuned using ratings from low variance raters for the case study subsections across the **(Left)** sleep and **(Right)** fitness domains. "$*$" indicates a statistically significant difference between two response types after multiple hypothesis testing correction. Error bars represent 95% confidence intervals.

Table D.20: **Generic prompting format for AutoEval model fine-tuning and inference.** Here, the "{VERTICAL}" placeholder is either sleep or fitness and "{TAG_LABEL}" is the case study section being rated (e.g., "Etiology"). "{OBJECTIVE}", "{PRINCIPLE_CRITERIA}", and "{OPTION_N}" refer to the evaluation criteria for the given section and principle as outlined in Appendix C.5. The "{RATING}" and "{RATING_TEXT_DESC}" placeholders are replaced with the rating assigned by the expert raters. Finally, "{CASE_STUDY_DATA}" is replaced with a high level overview of the case study while "{ASSISTANT_TEXT}" is replaced with the case study response from an expert or model.

Generic prompting format for AutoEval model fine-tuning and inference

**Prompt:**
You are a {VERTICAL} expert rater assessing the quality of health suggestions made by another personal health expert using the criteria below.

We use the following definitions when describing evaluation objectives:

- Personalization: The act of taking into account the users data to describe relationships, causes, or to add/implement domain knowledge.
- Important Data: Useful for accomplishing the objective.
- Unimportant Data: Not useful for accomplishing the objective.
- Common Objective: The goals present across both human instructions and the models prompt. Goals only performed by the model or only performed by the human rater are not included.

You are rating the quality of a {VERTICAL} "{TAG_LABEL}" response. The objective is: "{OBJECTIVE}".

The other {VERTICAL} expert was provided the following user data and information:

```
{CASE_STUDY_DATA}
```

The response from the other {VERTICAL} expert is:

```
{ASSISTANT_TEXT}
```

Grade this response using the following criteria and Likert scale statements:

Criteria: {PRINCIPLE_CRITERIA}

1. {OPTION_1}
2. {OPTION_2}
3. {OPTION_3}
4. {OPTION_4}
5. {OPTION_5}

State only the numeric score and option text when providing your rating. The formatting of your response must match that of the Likert scale statement.

**Target:**
{RATING}. {RATING_TEXT_DESC}

Table D.21: **Prompting format for AutoEval model fine-tuning and inference in the sleep vertical for the "Recommendations" section and the "No incorrect domain knowledge" principle.** Here, "{CASE_STUDY_DATA}" is replaced with a high level overview of the case study while "{ASSISTANT_TEXT}" is replaced with the case study response from an expert or model.

Prompting format for AutoEval model fine-tuning and inference in the sleep vertical for the "Recommendations" section and the "No incorrect domain knowledge" principle

**Prompt:**
You are a sleep expert rater assessing the quality of health suggestions made by another personal health expert using the criteria below.

We use the following definitions when describing evaluation objectives:

- Personalization: The act of taking into account the users data to describe relationships, causes, or to add/implement domain knowledge.
- Important Data: Useful for accomplishing the objective.
- Unimportant Data: Not useful for accomplishing the objective.
- Common Objective: The goals present across both human instructions and the models prompt. Goals only performed by the model or only performed by the human rater are not included.

You are rating the quality of a sleep "Recommendations" response. The objective is: "Provide recommendations to the user that can help them improve their sleep by addressing potential causes identified in the Etiology section. Avoid providing generic recommendations that are not personalized. This section does not require specific data to be cited directly, but the interpretation used to justify the recommendation should be present.".

The other sleep expert was provided the following user data and information:

```
{CASE_STUDY_DATA}
```

The response from the other sleep expert is:

```
{ASSISTANT_TEXT}
```

Grade this response using the following criteria and Likert scale statements:

Criteria: This section does not contain evidence of incorrect domain knowledge (e.g., factually incorrect or not accepted by expert consensus).

1. Only incorrect domain knowledge is referenced.
2. Many incorrect domain knowledge references exist.
3. Several incorrect domain knowledge references exist.
4. A few incorrect domain knowledge references exist.
5. No incorrect domain knowledge references exist.

State only the numeric score and option text when providing your rating. The formatting of your response must match that of the Likert scale statement.

**Target:**
5. No incorrect domain knowledge references exist.

Table D.22: **AutoEval model performance in the validation set across verticals.** AutoEval model rating predictions are compared with ground truth human ratings from the validation dataset. Here, "Gemini Pro" denotes an untuned baseline, "Primary" denotes models tuned on only one expert's ratings, and "All" denotes models tuned on all ratings for the given vertical. Spearman's rank correlation, Kendall's Coefficient of Concordance (Kendall's W), and Weighted Cohen's Kappa measurements were computed using all ratings for each principle and section. Mean metrics and 95% confidence intervals derived from 1,000 bootstrapping iterations are reported for each pair. Using paired bootstrapping, we find that all tuned AutoEval models significantly outperform the untuned baseline across metrics. However, due to low sample size, differences between tuned AutoEval models are not statistically significant.

| AutoEval Model | Spearman's R | Kendall's W | Weighted Kappa |
|---|---|---|---|
| **Fitness** | | | |
| Gemini Pro | 0.205 (0.152–0.256) | 0.198 (0.147–0.248) | 0.203 (0.152–0.256) |
| All | 0.280 (0.228–0.329) | 0.274 (0.224–0.322) | 0.152 (0.114–0.195) |
| Fitness Primary B | 0.284 (0.230–0.336) | 0.278 (0.225–0.329) | 0.142 (0.106–0.182) |
| Fitness Primary C | 0.305 (0.256–0.352) | 0.291 (0.245–0.335) | 0.320 (0.270–0.369) |
| **Sleep** | | | |
| Gemini Pro | 0.242 (0.185–0.299) | 0.230 (0.177–0.283) | 0.223 (0.166–0.280) |
| Sleep Primary D | 0.333 (0.279–0.384) | 0.316 (0.265–0.365) | 0.321 (0.256–0.389) |
| All | 0.341 (0.288–0.395) | 0.325 (0.273–0.378) | 0.347 (0.272–0.424) |
| Sleep Primary C | 0.368 (0.320–0.417) | 0.343 (0.298–0.388) | 0.389 (0.338–0.443) |

Table D.23: **Average time (m) taken to rate all responses for a single case across all sections and principles**. We consider primary raters, secondary raters, a single model, and a fully parallelized set of models, where "fully parallelized" means serving a model replica for each combination of sections and principles. The mean time-to-rate and the corresponding 95% confidence intervals were calculated across 1,000 bootstrapping iterations.

| Rater | Fitness | Sleep |
|---|---|---|
| Primary raters | 44.7 (38.6–51.2) | 26.8 (25.2–28.6) |
| Secondary raters | 87.7 (62.4–115.9) | 45.7 (34.3–60.9) |
| Model (single) | 27.5 (26.9–28.1) | 24.7 (23.7–25.6) |
| Model (Fully parallelized) | 0.367 (0.361–0.373) | 0.411 (0.402–0.421) |

# E Professional Examinations

## E.1 Additional ablation experiments

Table E.1: **Effects of chain-of-thought prompting and self-consistency on PH-LLM Accuracy for Sleep and Fitness Professional Exams.** CoT=Chain-of-Thought, SC=Self-Consistency.

| Domain | Prompt | SC | No SC |
|---|---|---|---|
| Sleep | CoT | **79%** | 78% |
| | No CoT | **79%** | **79%** |
| Fitness | CoT | **88%** | 84% |
| | No CoT | **87%** | 85% |

Table E.2: **Effects of Few-Shot prompting on PH-LLM Accuracy for Sleep and Fitness Professional Exams.**

| Domain | Prompt | SC + CoT |
|---|---|---|
| Sleep | Few-Shot | **79%** |
| | Zero-Shot | 75% |
| Fitness | Few-Shot | **88%** |
| | Zero-Shot | 87% |

## E.2 Comparison of Professional Exam Performance on Additional Models.

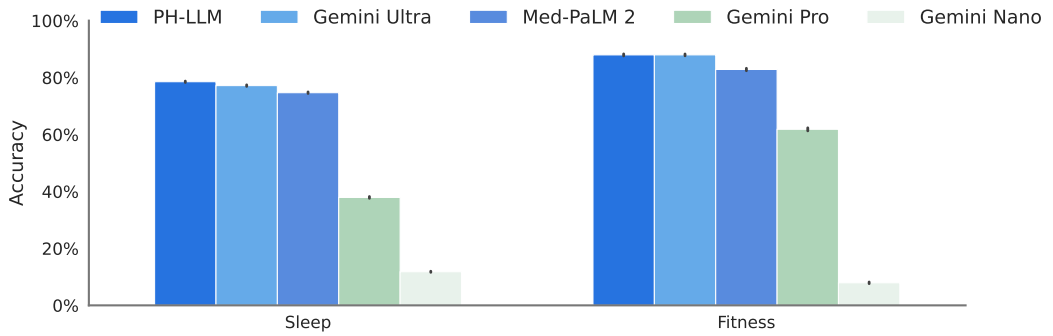



Figure E.1: **Overall performance of professional exams across PH-LLM, different Gemini models, and Med-PaLM 2.** All Gemini model sizes are based on the Gemini 1.0 model family. Error bars represent 95% confidence intervals.

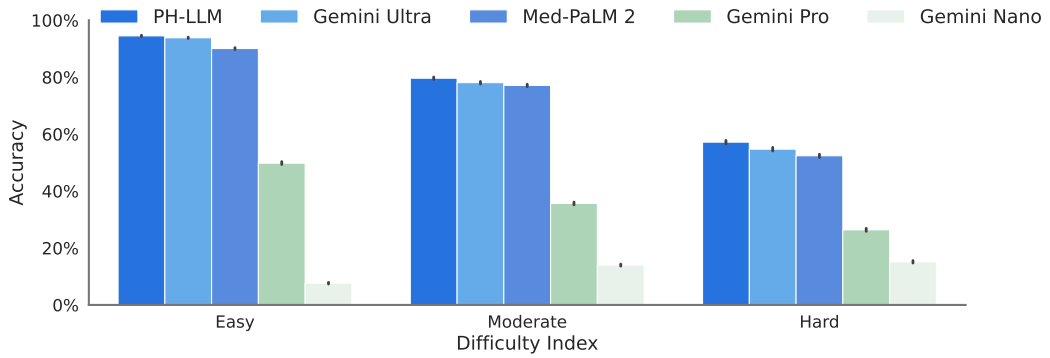



Figure E.2: **Breakdown performance of sleep professional exams across PH-LLM, different Gemini models, and Med-PaLM 2.** All Gemini model sizes are based on the Gemini 1.0 model family. Error bars represent 95% confidence intervals.

### E.3 Prompts Used for Professional Exams

Table E.3: **Prompt for Multiple-Choice Questions with Chain-of-Thought**

Prompt for Multiple-Choice Questions with Chain-of-Thought

**Question:**
Instructions: The following are multiple choice questions about {domain} knowledge. Solve them in a step-by-step fashion, starting by summarizing the available information. Output a single option from {mcq_options} as the final answer and enclosed by xml tags <answer></answer>.
Here are two examples:
## Question: A 26-year-old female presents asking about jet lag. She has no past medical history, lives on the East Coast, and travels frequently to the West Coast for business. The person's career involves planning evening events, and she reports significant sleepiness at these events that impairs her ability to perform her job. She wants to know how she can adapt to Pacific Standard Time (PST) before she travels. What treatment plan will help this patient adapt to PST prior to travel?
(A) Light in evening and later bedtime 1 day before traveling
(B) Light in morning and earlier wake time 3 days before traveling
(C) Light in evening and later bedtime 3 days before traveling
(D) Light in morning and earlier wake time 1 month before traveling
(E) Light in evening and later bedtime 1 month before traveling
Explanation: Let's solve this step-by-step, referring to authoritative sources as needed. The West Coast is 3 timezones behind the East Coast. Since she plans evening events, she needs to shift her schedule to stay up 3 hours later. Adding light in the evening will disrupt melatonin production, delaying sleepiness. Transitioning timezones typically takes one day per timezone.
Answer: <answer>(C)</answer>
## Question: What is a difference in the clinical features of obstructive sleep apnea (OSA) in older adults compared to younger adults?
(A) Increased prevalence of OSA among older adults occurs after age 65.
(B) Clinical symptoms associated with OSA (e.g. excessive daytime sleepiness) are less common and less severe in older adults than in younger adults.
(C) Increased risk of cardiopulmonary diseases is greater among elderly than among younger individuals.
(D) Excess body weight, snoring, and witnessed apneas more consistently indicate OSA in older adults than in younger individuals.
(E) There are no significant OSA differences between older and younger adults.
Explanation: Let's solve this step-by-step, referring to authoritative sources as needed. Compared to younger patients with the same apnea hypopnea index, OSA in older patients is associated with less sleepiness (Morrell et al 2012). This observation has led some to suggest that OSA in the elderly may represent a distinct physiological phenotype. Answer: <answer>(B)</answer>
## Question: {mcq_question}
Explanation: Let us solve this step-by-step, referring to authoritative sources as needed.

Table E.4: **Prompt for Multiple-Choice Questions without Chain-of-Thought**

Prompt for Multiple-Choice Questions without Chain-of-Thought

**Question:**
Instructions: The following are multiple choice questions about {domain} knowledge. Output a single option from {mcq_options} as the final answer and enclosed by xml tags <answer></answer>.
Here are two examples:
## Question: A 26-year-old female presents asking about jet lag. She has no past medical history, lives on the East Coast, and travels frequently to the West Coast for business. The person's career involves planning evening events, and she reports significant sleepiness at these events that impairs her ability to perform her job. She wants to know how she can adapt to Pacific Standard Time (PST) before she travels. What treatment plan will help this patient adapt to PST prior to travel?
(A) Light in evening and later bedtime 1 day before traveling
(B) Light in morning and earlier wake time 3 days before traveling
(C) Light in evening and later bedtime 3 days before traveling
(D) Light in morning and earlier wake time 1 month before traveling
(E) Light in evening and later bedtime 1 month before traveling
Answer: <answer>(C)</answer>
## Question: What is a difference in the clinical features of obstructive sleep apnea (OSA) in older adults compared to younger adults?
(A) Increased prevalence of OSA among older adults occurs after age 65.
(B) Clinical symptoms associated with OSA (e.g. excessive daytime sleepiness) are less common and less severe in older adults than in younger adults.
(C) Increased risk of cardiopulmonary diseases is greater among elderly than among younger individuals.
(D) Excess body weight, snoring, and witnessed apneas more consistently indicate OSA in older adults than in younger individuals.
(E) There are no significant OSA differences between older and younger adults.
Answer: <answer>(B)</answer>
## Question: {mcq_question}

Table E.5: **Prompt for Multiple-Choice Questions with Chain-of-Thought and Zero-Shot**

Prompt for Multiple-Choice Questions with Chain-of-Thought and Zero-Shot

**Question:**
Instructions: The following are multiple choice questions about {domain} knowledge. Solve them in a step-by-step fashion, starting by summarizing the available information. Output a single option from {mcq_options} as the final answer and enclosed by xml tags <answer></answer> such as <answer>(A)</answer>.
## Question: {mcq_question}

# F Patient-Reported Outcomes

## F.1 Prepossessing

While most individuals had sensor data for over 21 days, distributions were heavily left-skewed (Supplementary Figure F.1). To obtain a rectangular dataset we retained only individuals with at least 15 days of sensor data (N=7,114) and downsampled all individuals to a set of 15 contiguous days. We imputed all remaining missing values with the population median computed using all available data from training set individuals, resulting in a $20 \times 15$ matrix that represents the wearable sensor data for each research participant over 15 days. Furthermore, we performed standard filtering for data quality by removing any data points that were more than four standard deviations from the population median for each sensor value. No imputation was performed for survey answers.

For training, validation and final evaluation we split the dataset into three groups randomly resulting in 4,978 training examples, 703 validation examples and 1,433 examples.

## F.2 Patient-reported outcome prediction input features

Table F.1: **Sensor features used to predict each patient-reported outcome.**

| Sensor Feature | Definition |
|---|---|
| Heart rate variability (rmssd) | Heart rate variability root mean square of successive differences |
| Respiratory rate (rate_brpm) | Respiratory rate breaths per minute |
| Resting heart rate (rhr_bpm) | Resting heart rate beats per minute |
| Awake minutes (awake_minutes) | Awake minutes |
| Deep sleep minutes (deep_sleep_minutes) | Deep sleep minutes |
| Sleep duration (duration_minutes) | Sleep duration minutes |
| Sleep efficiency (efficiency) | The fraction of time in bed that was spent sleeping |
| Overall sleep score (overall_score) | Overall sleep score |
| Percent of sleep in REM (rem_sleep_percent) | Percent of sleep in REM |
| Restlessness (restlessness) | Restlessness |
| Revitalization score (revitalization_score) | Revitalization score |
| Sleep end time (sleep_end_time) | Sleep end time encoded as minutes after midnight |
| Sleep start time (sleep_start_time) | Sleep start time encoded as minutes after midnight |
| Sleep time (sleep_time_minutes) | Sleep time minutes |
| Awake state minutes (waso_count_long_wakes) | Total number of minutes in awake state after sleep onset |
| Number steps (num_steps) | Number of steps walked during the day |
| Cardio minutes (cardio_minutes) | Number of minutes spent in cardio zone during the day |
| Fat burn minutes (fat_burn_minutes) | Number of minutes spent in fat burn zone during the day |
| Peak minutes (peak_minutes) | Number of minutes spent in peak zone during the day |
| Total exercise time (total_multiplied_minutes) | Total multiplied minutes of exercise during the day |


Figure F.1: **Distribution of number of data points available for each sensor.**

### F.3 Patient-reported outcome surveys

Each survey is coded so that higher values correspond with greater sleep disturbance or impairment.

**Sleep Disturbance Survey**

In the past 7 days, my sleep was restless. [Very Restless]

| 1. Not at all | 2. A little bit | 3. Somewhat | 4. Quite a bit | 5. Very much |
|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ |

In the past 7 days, I was satisfied with my sleep. [Satisfied]

| 5. Not at all | 4. A little bit | 3. Somewhat | 2. Quite a bit | 1. Very much |
|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ |

In the past 7 days, my sleep was refreshing. [Refreshed]

| 5. Not at all | 4. A little bit | 3. Somewhat | 2. Quite a bit | 1. Very much |
|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ |

In the past 7 days, I had difficulty falling asleep. [Trouble falling asleep]

| 1. Not at all | 2. A little bit | 3. Somewhat | 4. Quite a bit | 5. Very much |
|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ |

In the past 7 days, I had trouble staying asleep. [Trouble staying asleep]

| 1. Never | 2. Rarely | 3. Sometimes | 4. Often | 5. Always |
|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ |

In the past 7 days, I had trouble sleeping. [Trouble sleeping]

| 1. Never | 2. Rarely | 3. Sometimes | 4. Often | 5. Always |
|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ |

In the past 7 days, I got enough sleep. [Enough sleep]

| 5. Never | 4. Rarely | 3. Sometimes | 2. Often | 1. Always |
|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ |

In the past 7 days, my sleep quality was. [Quality]

| 5. Very poor | 4. Poor | 3. Fair | 2. Good | 1. Very good |
|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ |

**Sleep Impairment Survey**

In the past 7 days, I had a hard time getting things done because I was sleepy. [Trouble being productive]

| 1. Not at all | 2. A little bit | 3. Somewhat | 4. Quite a bit | 5. Very much |
|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ |

In the past 7 days, I felt alert when I woke up. [Alert]

| 5. Not at all | 4. A little bit | 3. Somewhat | 2. Quite a bit | 1. Very much |
|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ |

In the past 7 days, I felt tired. [Tiredness]

| 1. Not at all | 2. A little bit | 3. Somewhat | 4. Quite a bit | 5. Very much |
|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ |

In the past 7 days, I had problems during the day because of poor sleep. [Having problems]

| 1. Not at all | 2. A little bit | 3. Somewhat | 4. Quite a bit | 5. Very much |
|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ |

In the past 7 days, I had a hard time concentrating because of poor sleep. [Sleep impairment due to trouble concentrating]

| 1. Not at all | 2. A little bit | 3. Somewhat | 4. Quite a bit | 5. Very much |
|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ |

In the past 7 days, I felt irritable because of poor sleep. [Sleep impairment due to irritability]

| 1. Not at all | 2. A little bit | 3. Somewhat | 4. Quite a bit | 5. Very much |
|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ |

In the past 7 days, I was sleepy during the daytime. [Sleepy during daytime]

| 1. Not at all | 2. A little bit | 3. Somewhat | 4. Quite a bit | 5. Very much |
|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ |

In the past 7 days, I had trouble staying awake during the day. [Trouble staying awake]

| 1. Not at all | 2. A little bit | 3. Somewhat | 4. Quite a bit | 5. Very much |
|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ |

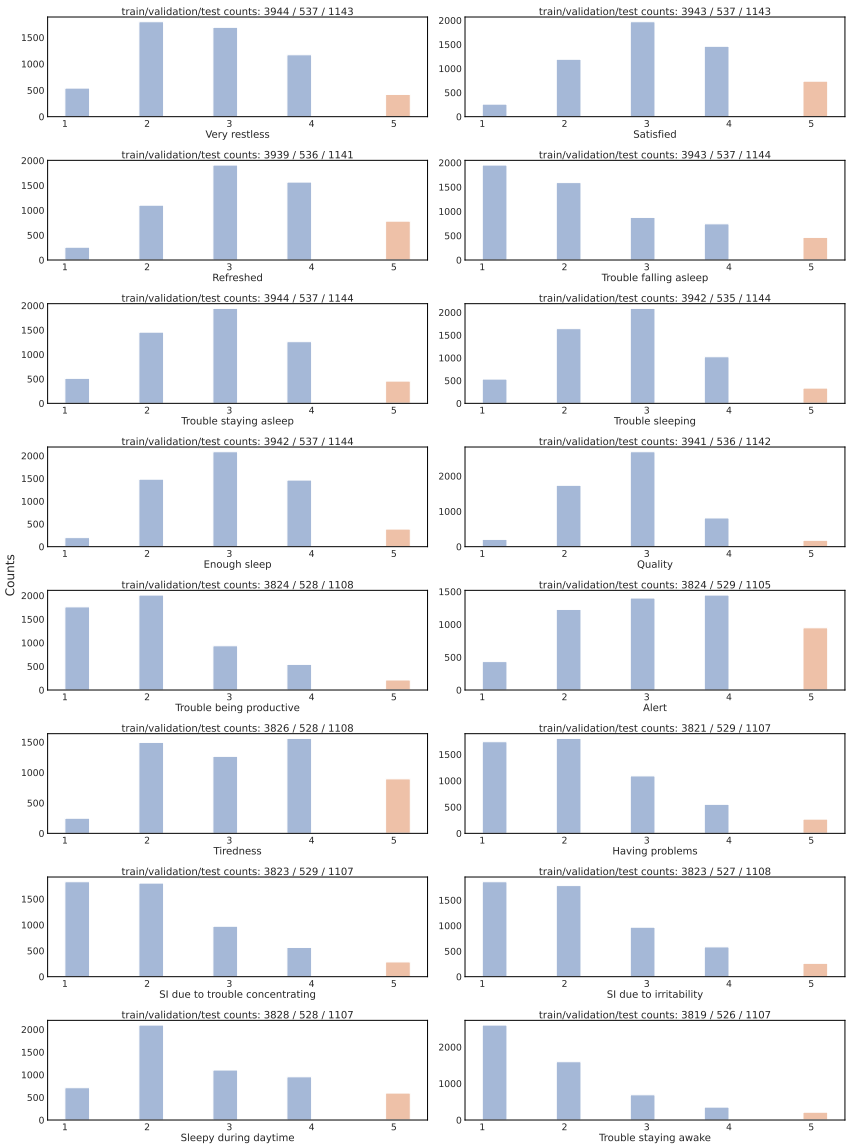
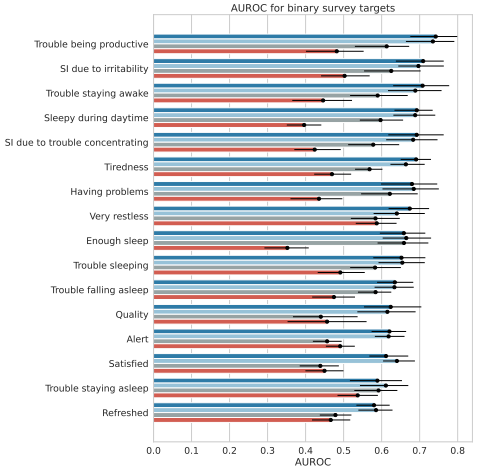
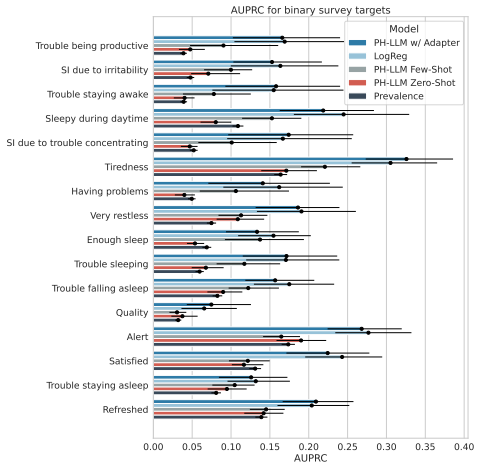



Figure F.2: **Distribution of responses for each survey question.** Survey questions are answered on a Likert scale. Here we show the distribution of responses for each question. The bar highlighted in a darker tone indicates those answers that were labeled as positive cases in the defined binary traits. The training, validation and test set counts are included in the title of each subplot.

### F.4 Patient-reported outcome prediction performance.




(a) Performance of PH-LLM measured via AUROC.




(b) Performance of PH-LLM measured via AUPRC.

Figure F.3: **Prediction of Patient-Reported Outcomes.** We evaluated the ability for the PH-LLM to infer subjective patient-reported outcomes using a multimodal adapter and compare to a suite of logistic regression models trained to predict each task independently, as well as PH-LLM using zero- and few-shot text prompting. Error bars represent 95% confidence intervals.

Table F.18: **Differences in AUROC between PH-LLM using a multimodal adapter and other modeling approaches.** Here, we highlight values in bold where the difference between PH-LLM with adapter and the other approach were **not** statistically significant. Statistical significance was determined via a paired bootstrap estimator.

| | PH-LLM zero-shot | PH-LLM few-shot | Logistic Regression |
|---|---|---|---|
| Very restless | 0.087 | 0.090 | **0.034** |
| Satisfied | 0.162 | 0.173 | **-0.029** |
| Refreshed | 0.114 | 0.101 | **-0.005** |
| Trouble falling asleep | 0.160 | **0.051** | **0.001** |
| Trouble staying asleep | **0.052** | **-0.003** | **-0.022** |
| Trouble sleeping | 0.161 | 0.069 | **-0.002** |
| Enough sleep | 0.306 | **-0.000** | **-0.006** |
| Quality | 0.167 | 0.184 | **0.009** |
| Trouble being productive | 0.260 | 0.129 | **0.007** |
| Alert | 0.130 | 0.164 | **0.002** |
| Tiredness | 0.221 | 0.122 | **0.027** |
| Having problems | 0.245 | **0.058** | **-0.005** |
| SI due to trouble concentrating | 0.268 | 0.114 | **0.009** |
| SI due to irritability | 0.207 | 0.084 | **0.012** |
| Sleepy during daytime | 0.296 | 0.095 | **0.004** |
| Trouble staying awake | 0.262 | 0.118 | **0.019** |

Table F.19: **Differences in AUPRC between PH-LLM using a multimodal adapter and other modeling approaches.** Here, we highlight values in bold where the difference between PH-LLM with adapter and the other approach were **not** statistically significant. Statistical significance was determined via a paired bootstrap estimator.

| | PH-LLM zero-shot | PH-LLM few-shot | Logistic Regression |
|---|---|---|---|
| Very restless | 0.077 | 0.073 | **-0.004** |
| Satisfied | 0.108 | 0.103 | **-0.019** |
| Refreshed | 0.067 | 0.064 | **0.005** |
| Trouble falling asleep | 0.067 | **0.035** | **-0.018** |
| Trouble staying asleep | **0.031** | **0.021** | **-0.006** |
| Trouble sleeping | 0.104 | 0.054 | **0.001** |
| Enough sleep | 0.080 | **-0.004** | **-0.021** |
| Quality | 0.037 | 0.044 | **0.009** |
| Trouble being productive | 0.118 | 0.076 | **-0.003** |
| Alert | 0.078 | 0.103 | **-0.009** |
| Tiredness | 0.155 | 0.105 | **0.021** |
| Having problems | 0.101 | 0.035 | **-0.021** |
| SI due to trouble concentrating | 0.127 | 0.073 | **0.008** |
| SI due to irritability | 0.082 | 0.053 | **-0.011** |
| Sleepy during daytime | 0.138 | 0.066 | **-0.026** |
| Trouble staying awake | 0.118 | 0.080 | **0.003** |

### F.5 Patient-reported outcome prompt examples

Table F.20: **Example of prompt given to PH-LLM to score PROs.** Demographic and sensor values are passed as text to the model. The feature to predict (in this example, `very restless`) can then be scored using the potential completions "yes." or "no." For few-shot prompting we additionally prepend complete examples from the training set to the prompt. When using the multimodal adapter (see Methods), a vector representation of the quantitative data is passed in via a set of learned tokens. Values in the below prompt are synthetic.

Example of prompt given to PH-LLM to score PROs.

Use the information provided to predict "very restless".
age: [40-45]. heart rate variability root mean square of successive differences: 33.5. respiratory rate breaths per minute: 16.5. resting heart rate beats per minute: 60.0. awake minutes: 51.0. deep sleep minutes: 53.0. sleep duration minutes: 471610.0. efficiency: 0.85. overall sleep score: 81.0. percent of sleep in REM: 16.0. restlessness: 0.07. revitalization score: 83.2. sleep end time: -274.0. sleep start time: 364.0. sleep time minutes: 420.8. total number of minutes in awake state after sleep onset: 7.4. number of steps walked during the day: 6850.0. number of minutes spent in cardio zone during the day: 6.7. number of minutes spent in fat burn zone during the day: 18.9. number of minutes spent in peak zone during the day: 0.41. total multiplied minutes of exercise during the day: 45.32.
very restless: yes or no?

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The introduction section has a bullet list of contributions, which are summarized in the abstract as well. The list of claims/contributions matches the experimental results shown in figures throughout the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations can be found in Appendix B.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA] .

   Justification: There are no theory or proofs in this paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper specifies data collection details, training details, evaluation details, such as dataset splits and hyperparamters, wherever possible. Details pertaining to base Gemini model training are omitted as this code and training procedure is not open source.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The Gemini model weights are proprietary. The dataset is not open-sourced at this time. However, the methodological approach described in the paper can be replicated using an alternative large language model and relevant personal health data.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies training details, such as dataset splits and hyperparamters, wherever possible. Details pertaining to base Gemini model training are omitted as this code and training procedure is not open source.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All core results are accompanied by error bars, confidence intervals, and statistical significance tests. The factors of variability that the error bars are capturing are clearly stated, the method for calculating the error bars is explained, and any assumptions made are given.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: We are unable to release details on compute resources used to train or fine-tune Gemini models.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in this paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The broader impacts can be found in Appendix A.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No data or models are released as part of this manuscript.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All external data sources, such as multiple choice examinations, are described and credited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: This investigation performs retrospective selection and analysis of a de-identified dataset of individuals who gave optional specific informed consent and thus no specific instructions were given to participants. The general guidelines given to sleep and fitness experts are summarized in Section 2. The experts were paid at least the minimum wage (based on CA, USA).

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: This investigation performs retrospective selection and analysis of a de-identified dataset such that there is no potential harm or risk to study participants.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.