nature medicine

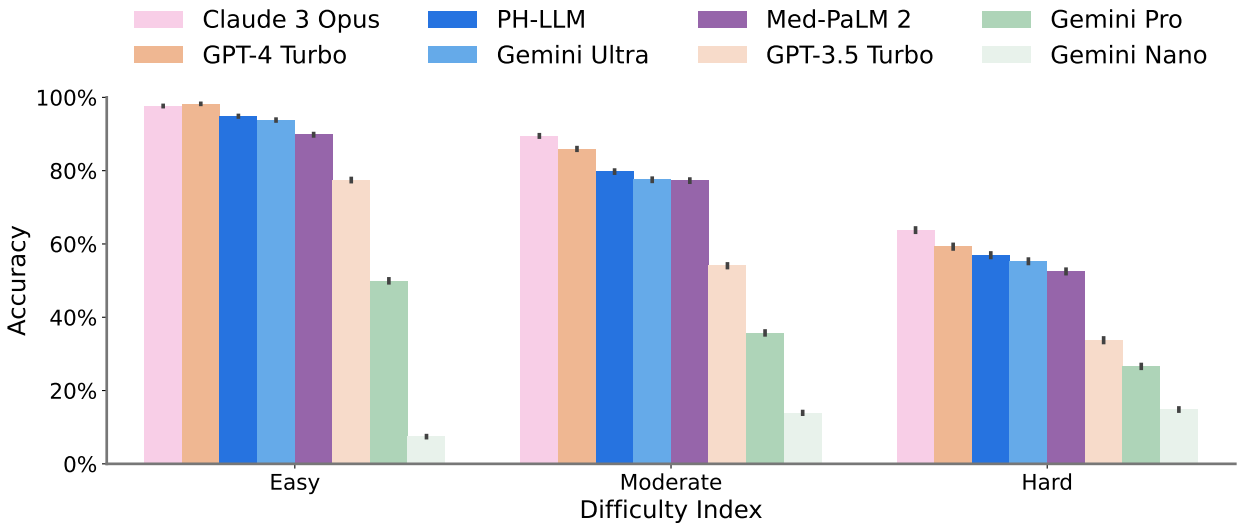
Supplementary information

https://doi.org/10.1038/s41591-025-03888-0

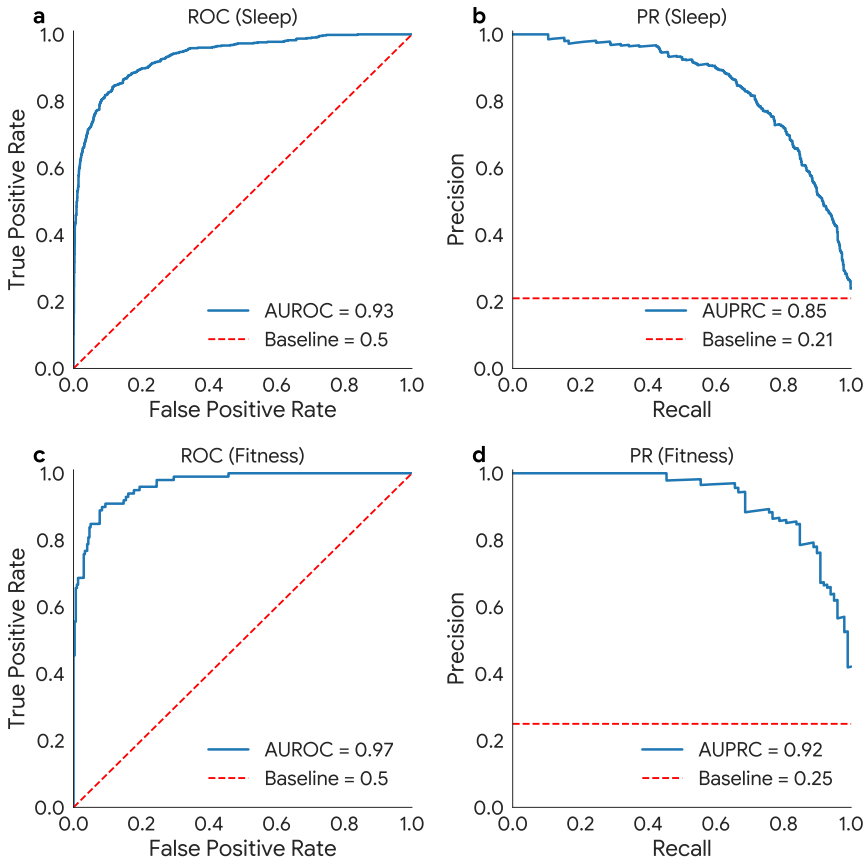
A personal health large language model for sleep and fitness coaching

In the format provided by the authors and unedited

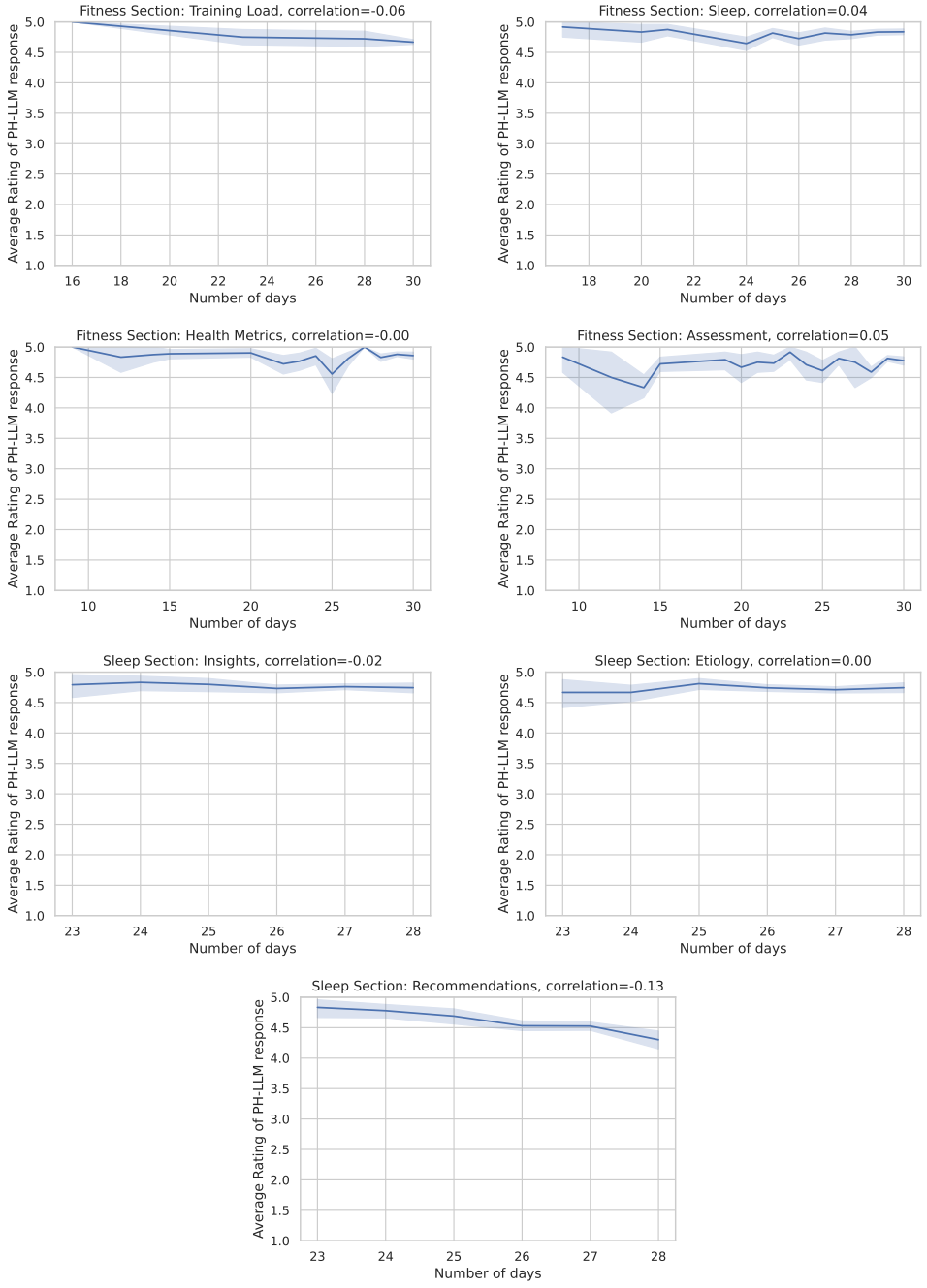
Supplementary Figures



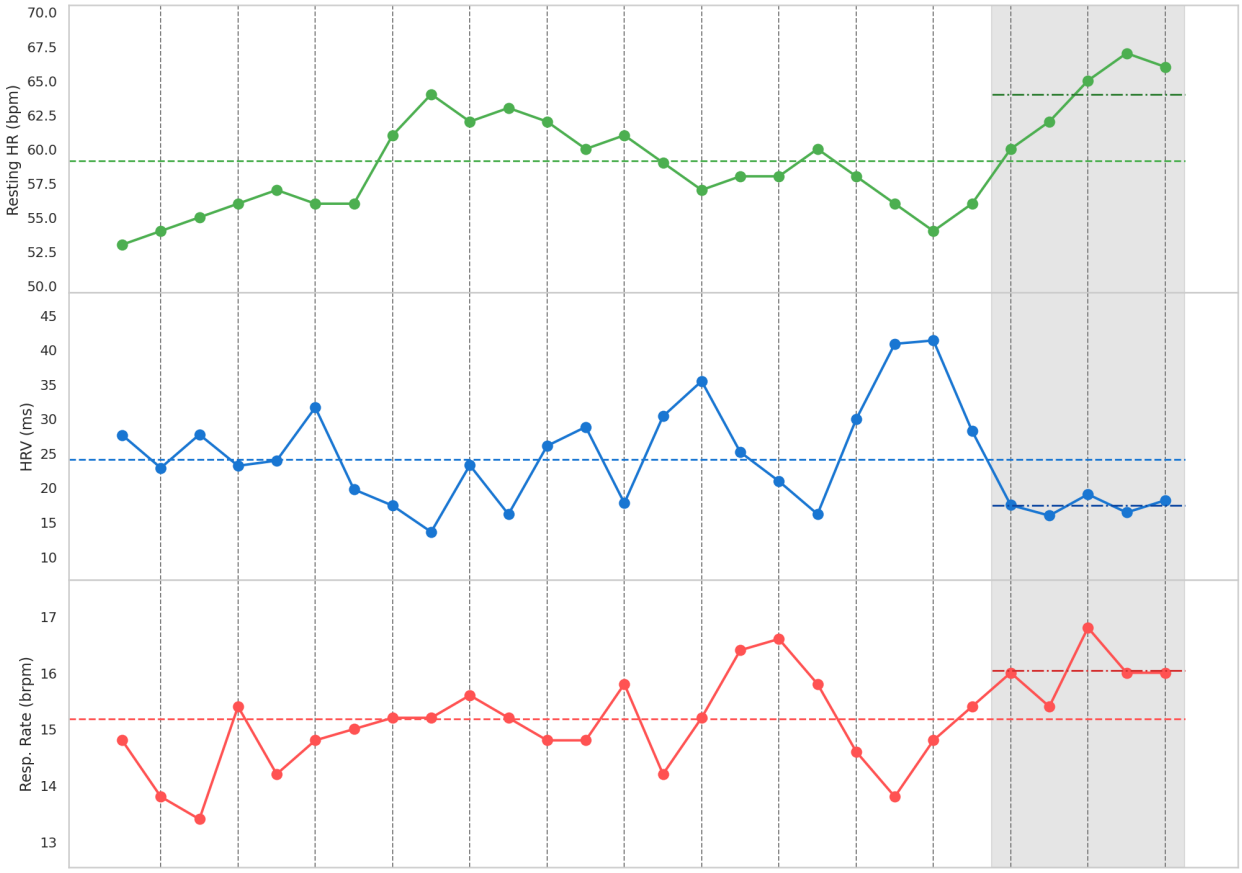
Supplementary Fig. 1 | Sleep professional exam performance stratified by question difficulty for PH-LLM, different Gemini models, GPT models, Claude 3 Opus, and Med-PaLM 2. All Gemini model sizes are based on the Gemini 1.0 model family. Error bars represent 95% confidence intervals derived from 1,000 bootstrapping iterations.



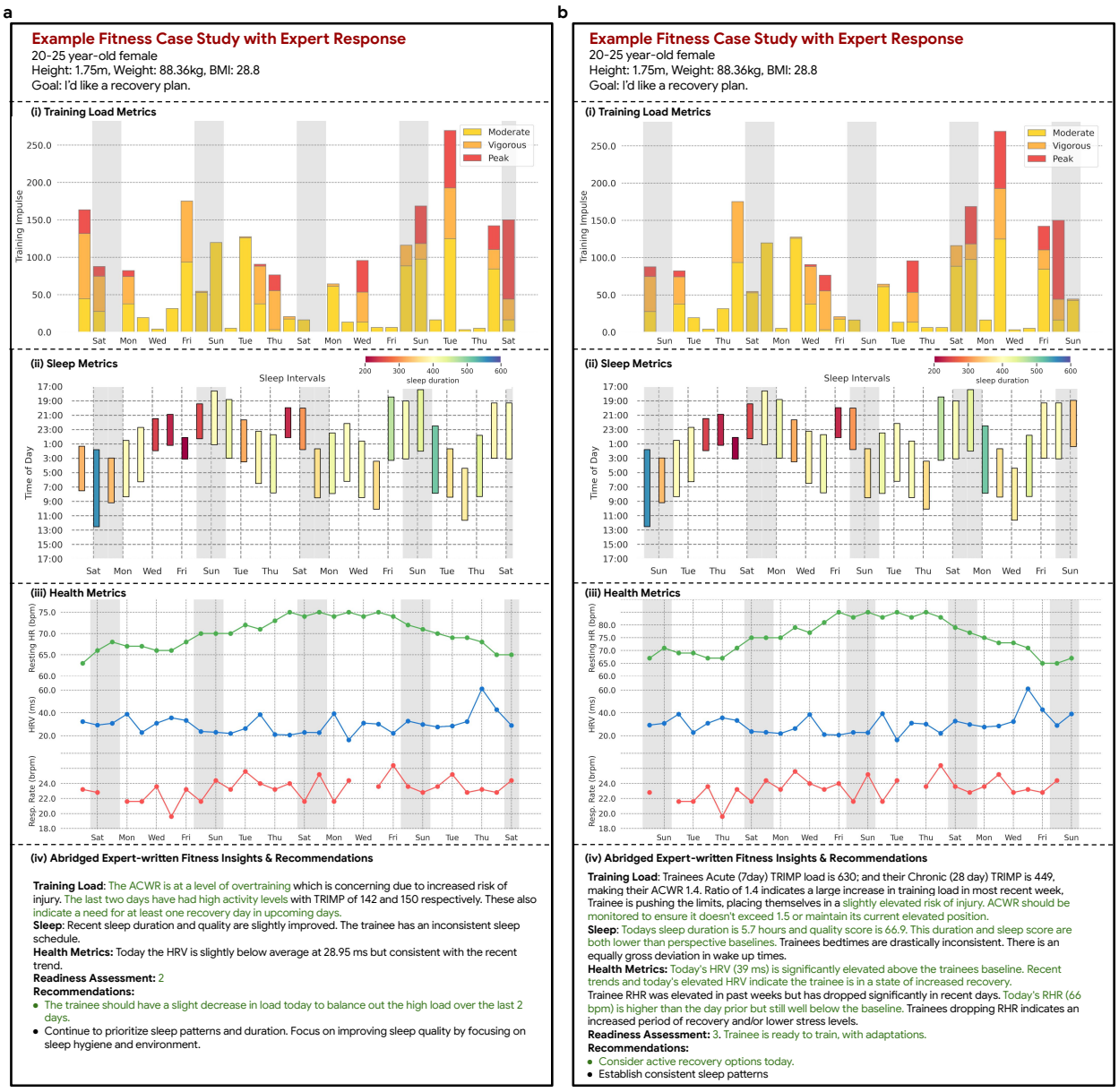
Supplementary Fig. 2 | Evaluation of PH-LLM answer confidence on professional exam questions. Receiver operating characteristic (ROC) curves (left) and precision-recall (PR) curves (right) computed using log probabilities assigned by PH-LLM to each answer choice. **a**, ROC curve for all sleep exam questions (n=629). **b**, PR curve for all sleep exam questions. **c**, ROC curve for all fitness exam questions (n=99). **d**, PR curve for all fitness exam questions. Area under the curves indicate micro AUC and AUPRC computed on softmaxed log probabilities. Dashed red lines indicate chance performance. Note the prevalence for sleep is 0.21 since it contains both 4-choice and 5-choice questions.



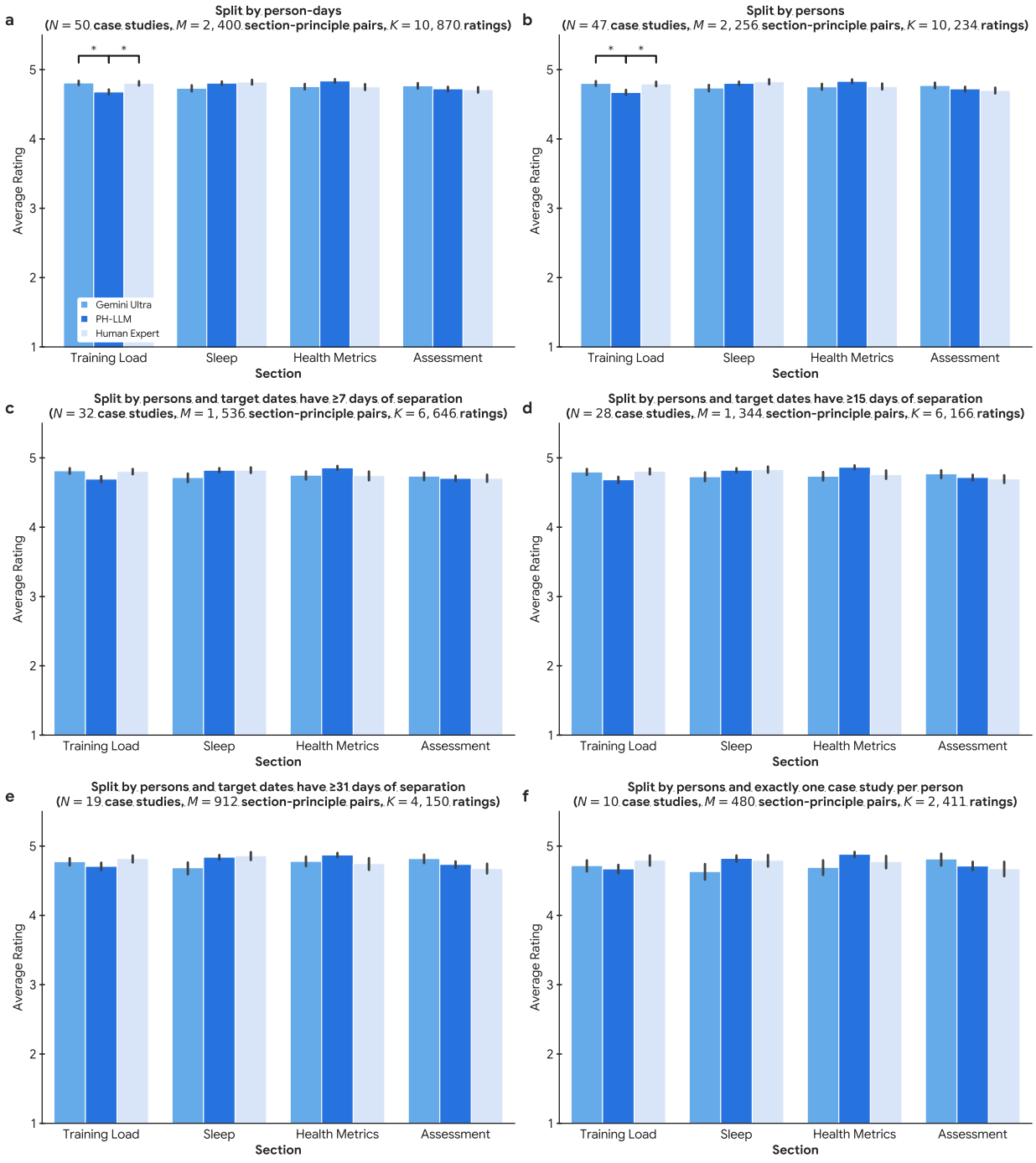
Supplementary Fig. 3 | Number of days of data available versus PH-LLM's performance on case studies. For coaching recommendations, the data naturally has varying lengths (e.g. due to individuals not using the Fitbit device some of the days). For each section in fitness and sleep case studies, we obtained the number of non-missing days and plot it against PH-LLM's performance as measured by the average expert rating (with 95% confidence interval) given to its response across the specific number of days available. We additionally include the correlation between number of days and the rating. For Fitness Section: Assessment, since it aggregates data across multiple sections, we determined the number of days by intersecting the non-missing days across training load, sleep, and health metrics sections and counting the number of days. We observed that the average rating does not correlate with the length of the input data. The consistent average rating across varying input lengths is expected since just like human experts, PH-LLM is limited by the information provided.



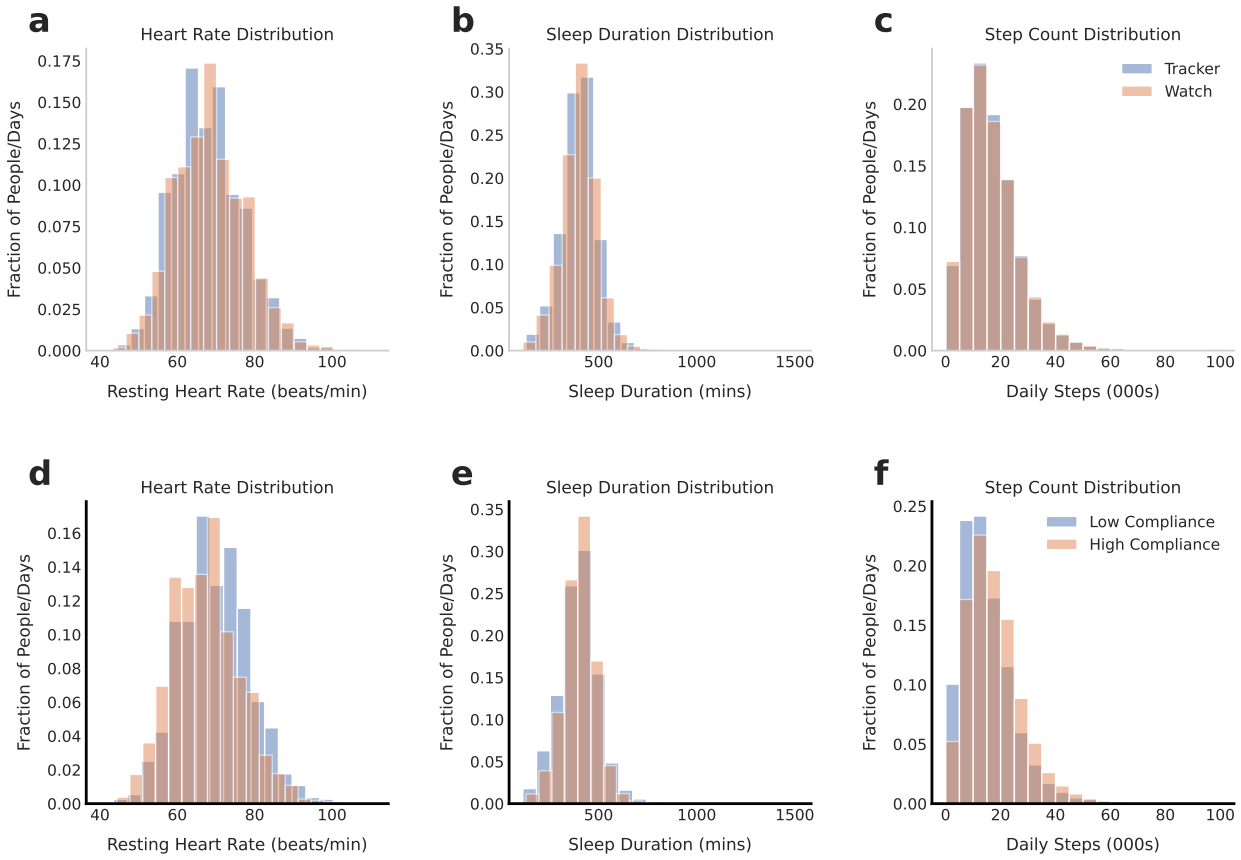
Supplementary Fig. 4 | Health metrics for all days with the last five days highlighted in gray. To demonstrate how length of input data affects conclusions, we analyzed how PH-LLM's response changes when using five days versus 30 days of data. For clarity, we highlight the five days that were analyzed along with changes in means for each metric. The dashed colored lines indicate the mean over all days and dashed-dotted darker colored lines indicate the mean computed over the last 5 days. The original PH-LLM response with 30 days is in Supplementary Table 16 and the updated PH-LLM response based on five days of data is in Supplementary Table 18.



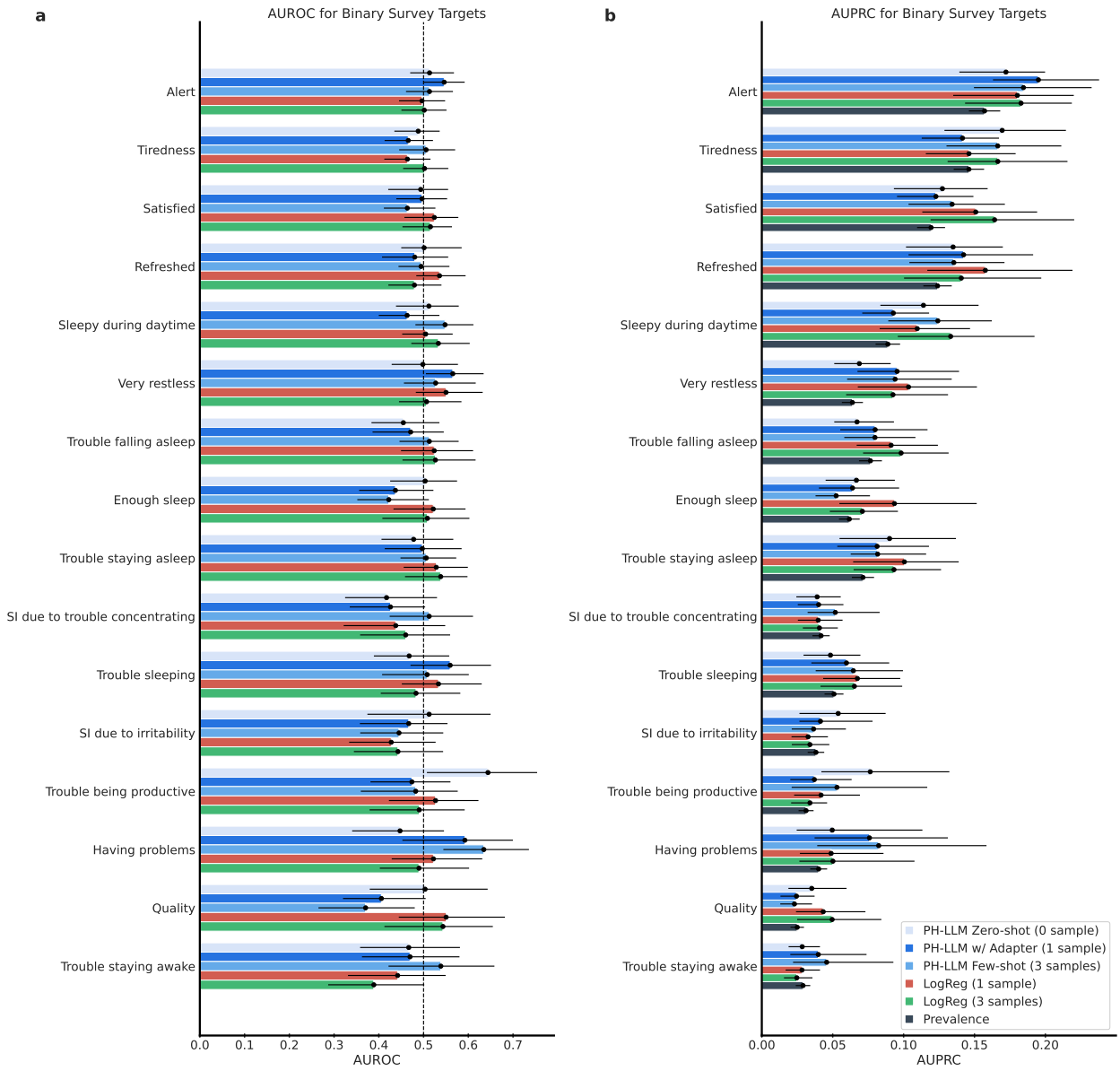
Supplementary Fig. 5 | Variation in expert assessment of fitness case studies from a single individual with one day offset. a, Base period. **b**, Period shifted forward by one day. In green, we highlight the sections of the expert assessment that are relevant to the shift. The presence of a training load rest day on last Sunday in **(b)** influences expert interpretation: while a rest day is recommended in **(a)**, in **(b)** the expert instead advises monitoring the high training load and increases the readiness score from 2 to 3 accordingly. Certain observations, such as irregular sleep patterns, remain consistent across both assessments.



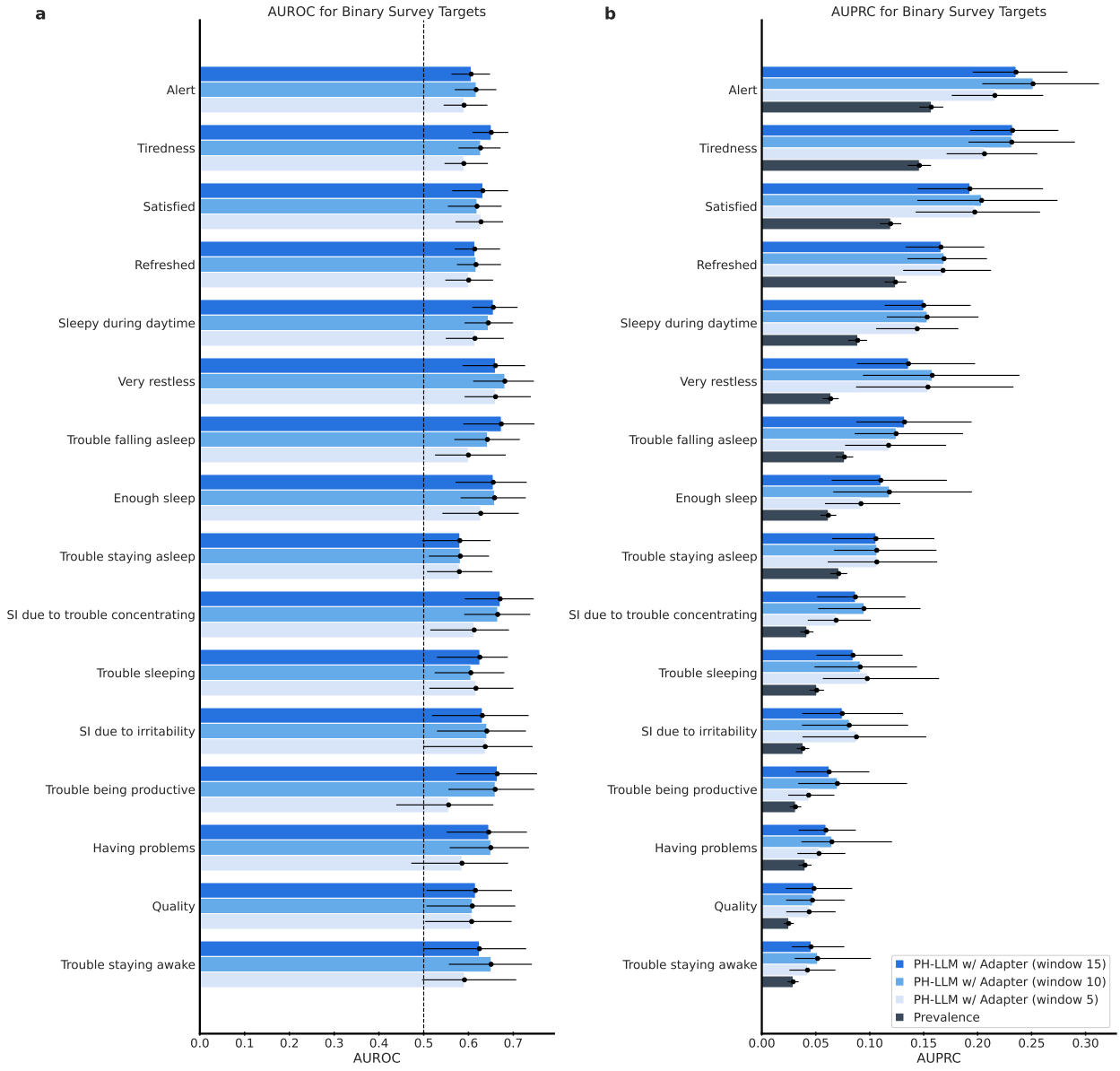
Supplementary Fig. 6 | Mean ratings given by experts for fitness case study subsections across holdout dataset partitioning methods. Across the train and holdout splits, person-months are unique though a single individual contributes case studies in both splits ($N = 5$ in the train set, $N = 3$ in the holdout set). Within the holdout split, some individuals contribute multiple case studies with overlapping days. **a**, The holdout dataset contains no person-days spanning splits (identical to Fig. 2d). **b**, The holdout dataset contains no persons from the train set. **c**, The holdout dataset contains no persons from the train set and, additionally, holdout case study target dates for each individual are separated by at least seven days. **d,e**, Analogous to (**c**) with target dates separated by at least 15 days (**d**) or 31 days (**e**). **f**, The holdout dataset contains exactly one case study per person. Though models are no longer statistically significantly different in the training load section due to decreased sample size, we observe that ranked preference ordering is consistent across methods.



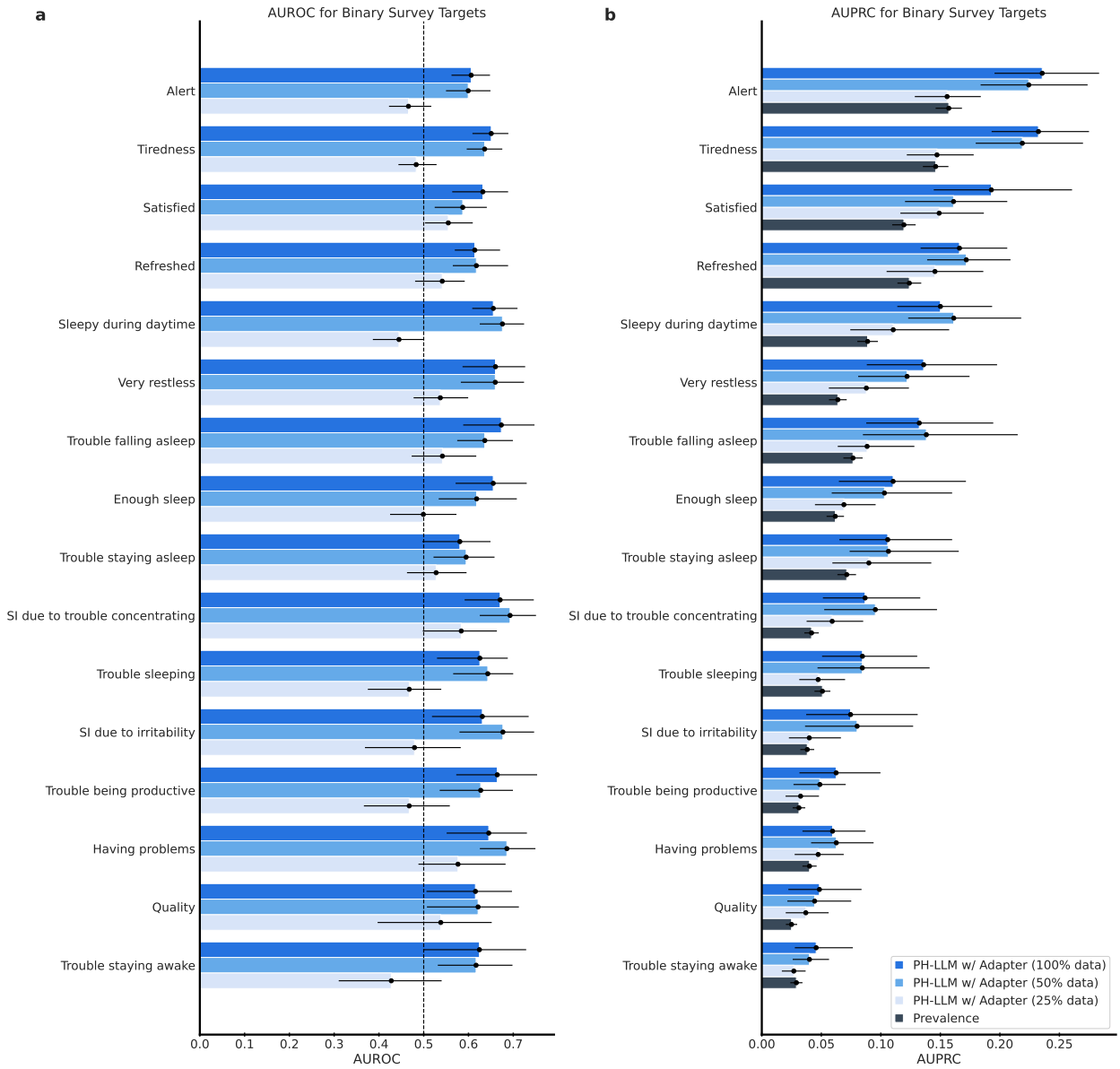
Supplementary Fig. 7 | Distributions of sensor data by device type and participant compliance. a-c, Resting heart rate, sleep duration, and step count distributions by device type (tracker vs watch). Trackers include Fitbit Charge, Fitbit Luxe, and Fitbit Inspire. Watches include Fitbit Sense and Google Pixel Watch. **d-f,** Resting heart rate, sleep duration, and step count distributions by participant compliance (low versus high). Low compliance participants were defined as having 20 or fewer days of sensor data.



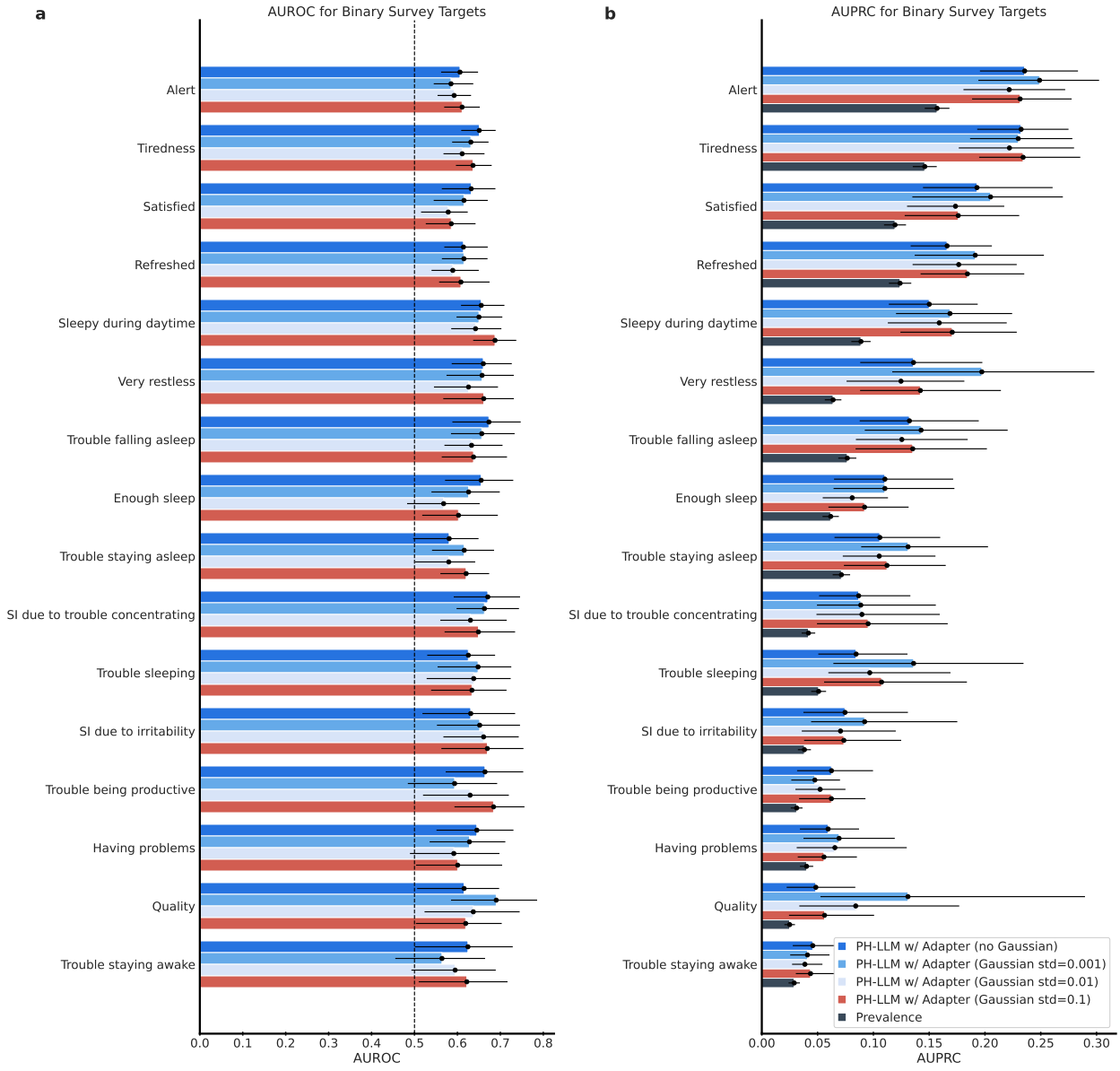
Supplementary Fig. 8 | PH-LLM performance on PRO prediction in the low-data regime. We evaluated the ability of PH-LLM with and without adapter, and logistic regression models to infer patient-reported outcomes when trained on 0-3 samples. **a,** Per-outcome AUROC. **b,** Per-outcome AUPRC. In the case of extremely limited amount of training data, PH-LLM (with and without an adapter) does not outperform logistic regression. This indicates that PH-LLM's pre-existing knowledge does not help interpret how wearable data relates to sleep outcomes, likely because its training data contains limited information about such relationships. This is expected given the scarcity of information about wearable data compared to, for example, general medical information. Error bars represent 95% confidence intervals.



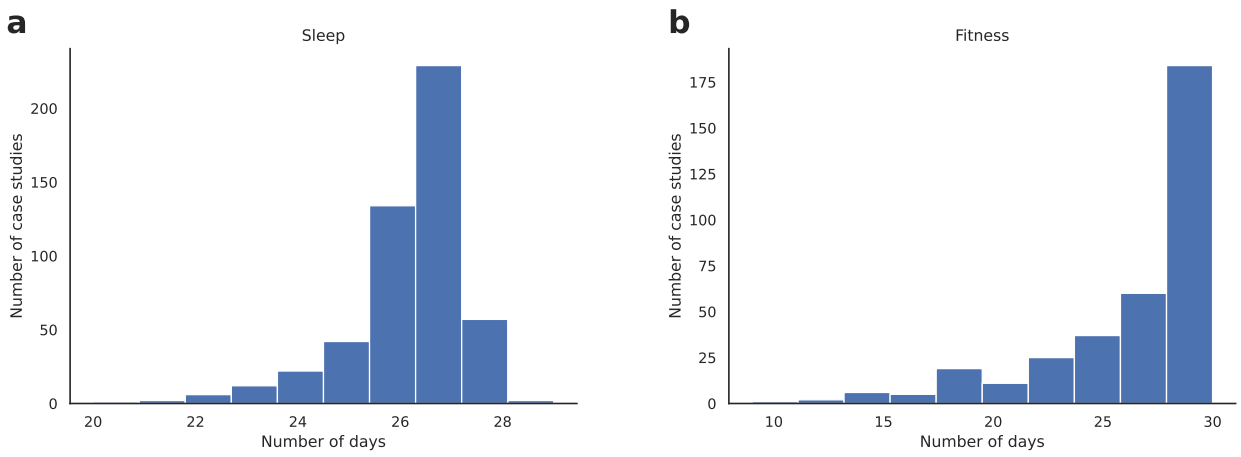
Supplementary Fig. 9 | Effect of sensor window size on PRO prediction. We evaluated the ability of PH-LLM with adapter to infer subjective patient-reported outcomes when the adapter inputs are aggregated from the last 15, 10, or five days of sensor data. Per-outcome AUROC (**a**) and AUPRC (**b**) with error bars and expected performance of the random predictor shown. Using the last 10 instead of five days of sensor data generally improves the performance as measured by AUROC and AUPRC while increasing from 10 to 15 days yields similar and sometimes worse performance. Error bars represent 95% confidence intervals.



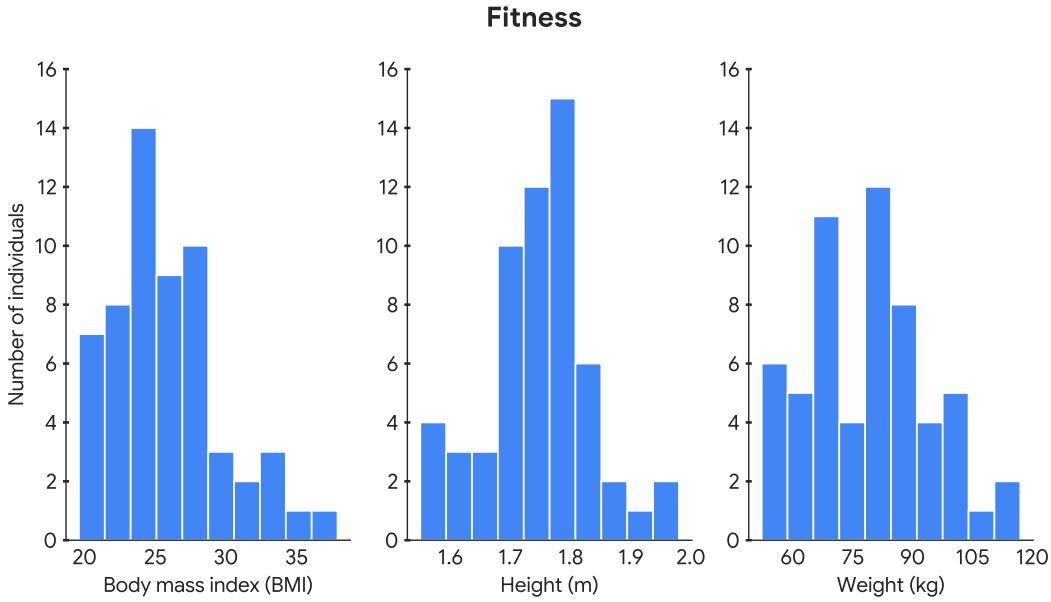
Supplementary Fig. 10 | Effect of training data size on PRO prediction. We evaluated the ability of PH-LLM with adapter to infer subjective patient-reported outcomes when provided with different amounts of training data. Per-outcome AUROC **(a)** and AUPRC **(b)** of PH-LLM with adapter trained with 2,914 (100%) samples, 1,457 (50%) samples, or 729 (25%) samples. Error bars represent 95% confidence intervals.



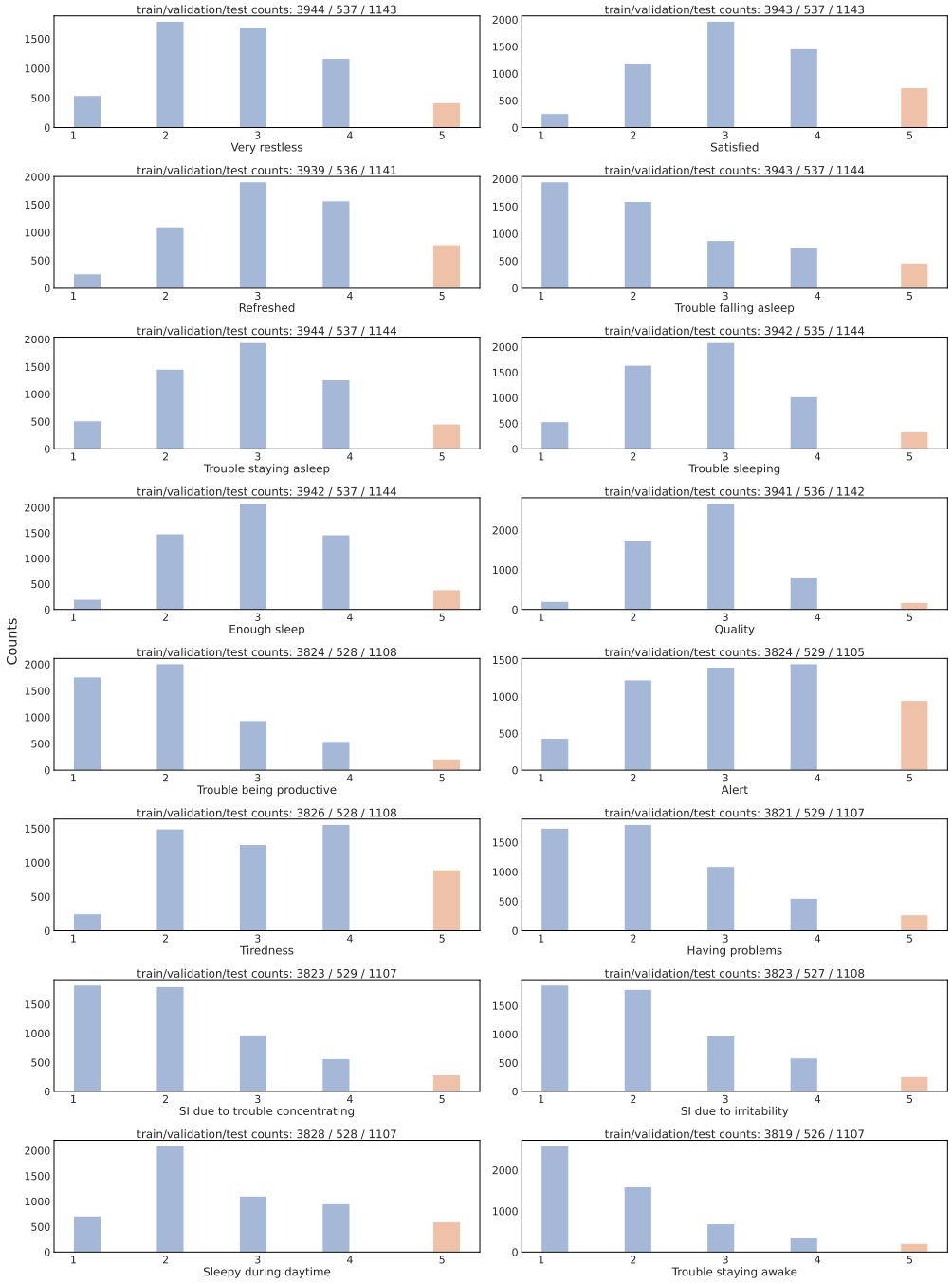
Supplementary Fig. 11 | Effect of data augmentation with Gaussian noise on PRO prediction. We evaluated the ability of PH-LLM with adapter to infer subjective patient-reported outcomes when the input in training data is augmented with Gaussian noise of 0 mean and different scale of standard deviation: 0 (no Gaussian noise), 0.1, 0.01, and 0.001. Error bars represent 95% confidence intervals. While the AUROC (**a**) scores do not show consistent trends, the AUPRC (**b**) scores are generally higher at the cost of larger error bars when Gaussian noise with zero mean and 0.001 standard deviation is added to each training data point independently.



Supplementary Fig. 12 | Histogram of the number of days of complete data available per case study. **a,** Sleep, **b,** Fitness case studies. Complete days were defined as those with daily metrics present for all sections of the case study.

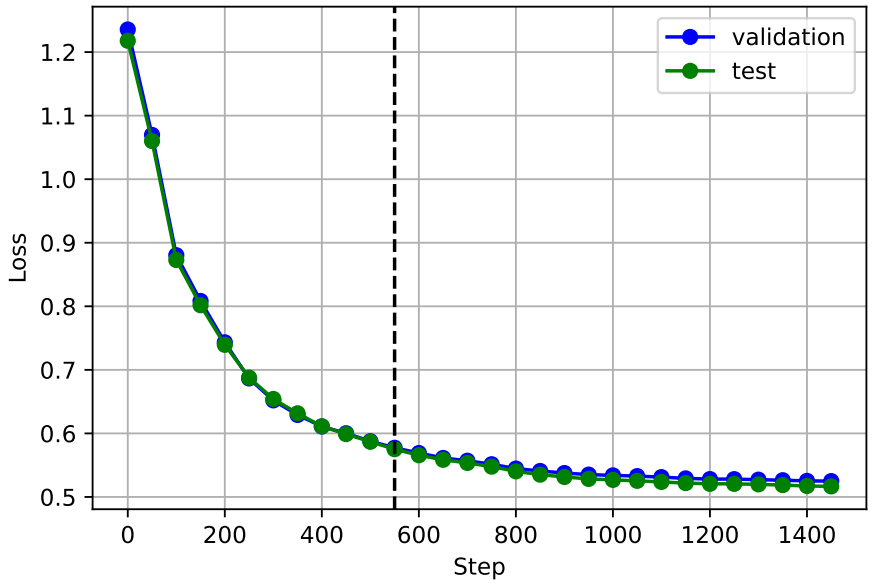


Supplementary Fig. 13 | Distributions of self-reported BMI, height, and weight in fitness case studies. Each of these features was collected and used for all individuals in fitness case studies. These features were not collected for sleep case studies because sample selection did not enrich for these variables and they were deemed unnecessary for assessing sleep patterns.

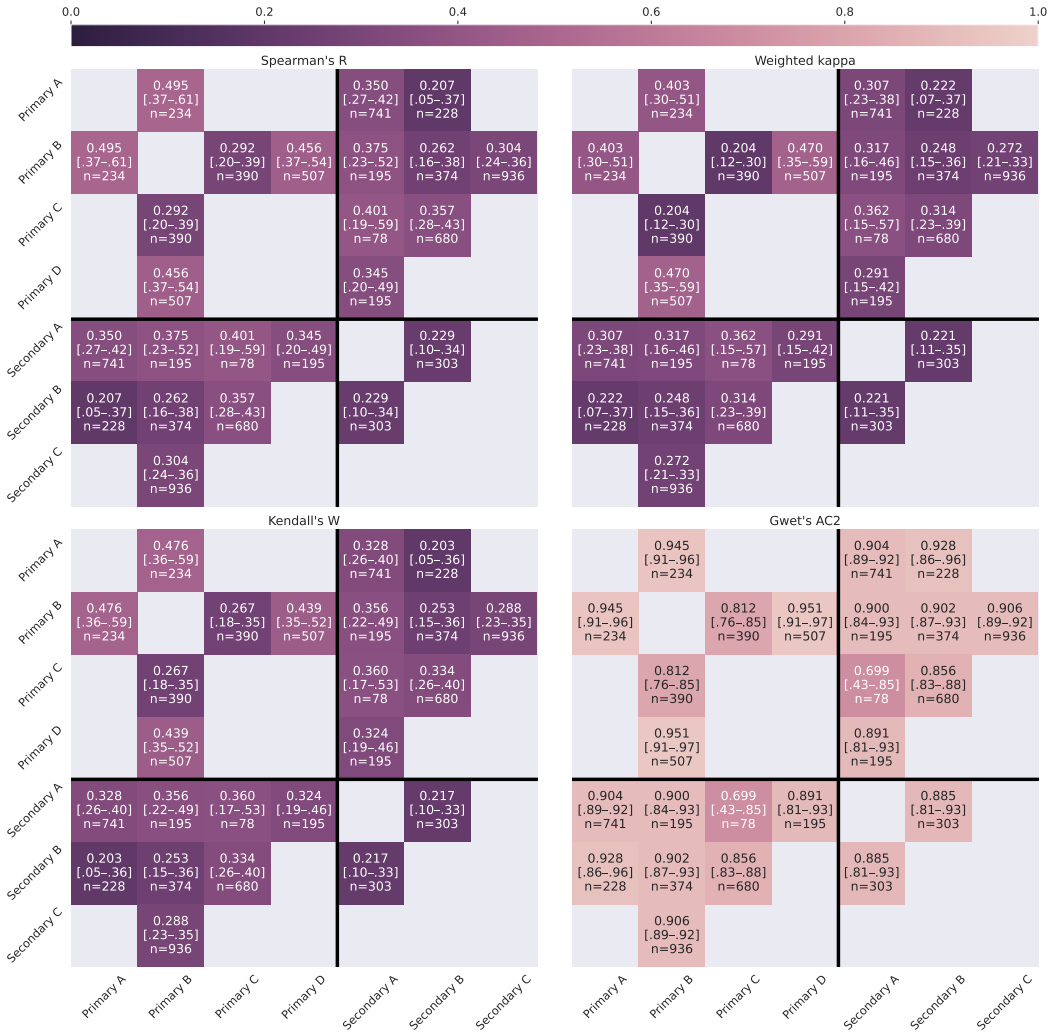


Supplementary Fig. 14 | Distribution of responses for each PRO survey question. Survey questions are answered on a 5-point Likert scale. Here we show the distribution of responses for each question. The bar highlighted in orange indicates those answers that were labeled as positive cases in the defined binary traits. The training, validation, and test set counts are included in the title of each subplot.

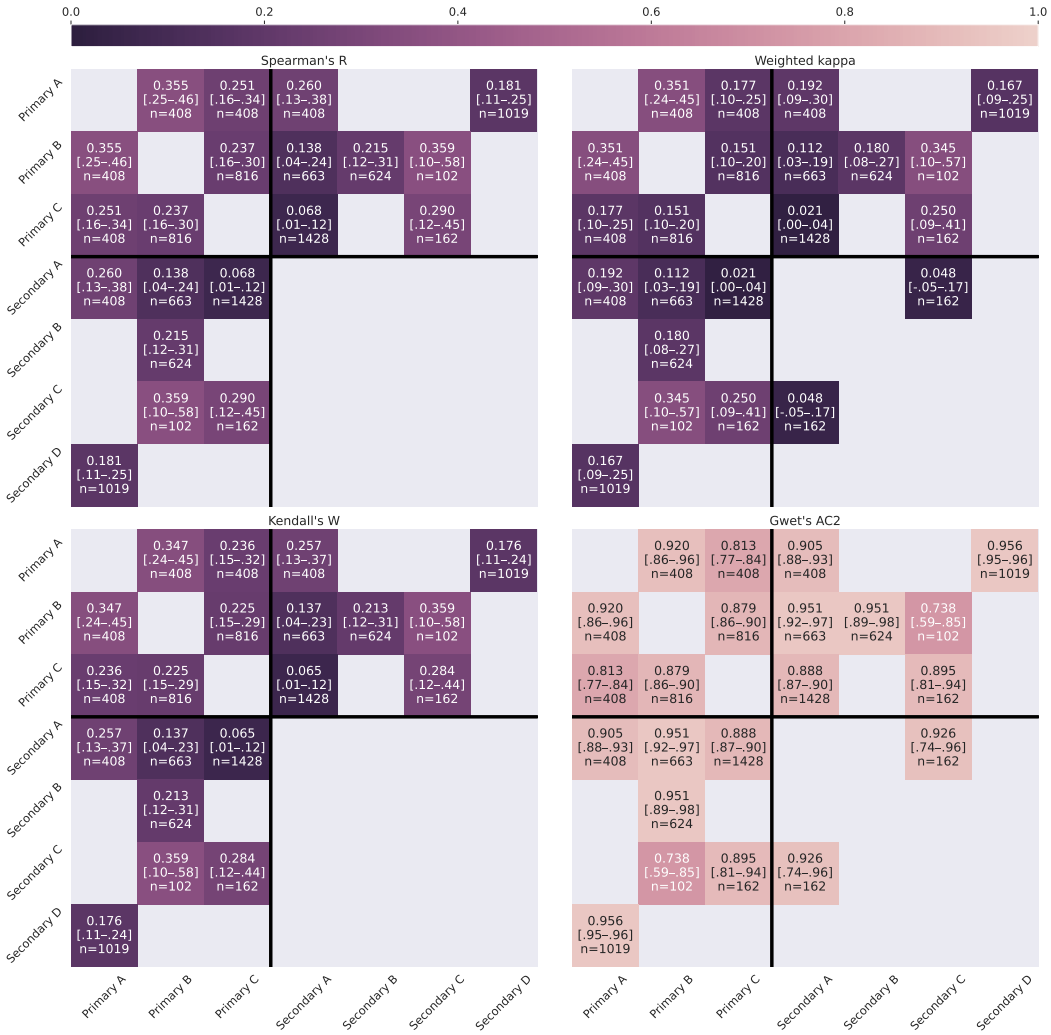
Supplementary Fig. 15 | Distribution of the number of data points available for each sensor in the patient-reported outcome dataset.



Supplementary Fig. 16 | Fine-tuning loss on validation and test set of case studies. The dashed line indicates the chosen checkpoint step for PH-LLM.



Supplementary Fig. 17 | 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.



Supplementary Fig. 18 | 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.

Supplementary Tables

Supplementary Table 1 | Accuracy of PH-LLM and Gemini Ultra 1.0 on MedQA and PubMedQA medical benchmarks. Chain-of-thought prompting was used to obtain results. Bold values indicate the highest performance.

Dataset	Gemini Ultra 1.0	PH-LLM
MedQA	**81%**	**81%**
PubMedQA	72%	**73%**

Supplementary Table 2 | 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.

Supplementary Table 3 | 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}

Supplementary Table 4 | 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}

Supplementary Table 5 | 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. Bold values indicate the highest performance.

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

Supplementary Table 6 | Effects of few-shot prompting on PH-LLM accuracy for sleep and fitness professional exams. SC + CoT=Self-Consistency and Chain-of-Thought prompting. Bold values indicate the highest performance.

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

Supplementary Table 7 | Examples of multiple-choice questions correctly answered by PH-LLM. Separate examples for sleep and fitness are shown. In green, we provide a comment from the sleep and fitness expert about the model response.

Examples of multiple-choice questions correctly answered by PH-LLM

Sleep Question:
A 26-year-old woman inquires about managing jet lag. She has no medical history and resides on the East Coast, frequently traveling to the West Coast for work. Her job involves organizing evening events, and she experiences considerable drowsiness during these events, hindering her job performance. She seeks advice on how to adjust to Pacific Standard Time (PST) before her trips. What strategy will assist her in acclimating to PST before she travels?
(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
Model Response:
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. Therefore, she needs to start the treatment 3 days before traveling. Answer: <answer>(C)</answer>.
Expert Comment: "The model identifies the differential in circadian phase (transitioning 3 time zones) and the needed direction of the phase shift (a delay), incorporates multiple essential concepts about circadian physiology (e.g., light's effect on melatonin based on timing), and establishes an appropriate plan for adjustment."

Fitness Question:
When a male gymnast performs a push-up, which muscle action occurs in the erector spinae muscle group?
(A) Concentric
(B) Eccentric
(C) Isometric
(D) Myocentric
Model Response:
The erector spinae muscle group is located in the back and helps to extend the spine. When a gymnast performs a push-up, the erector spinae muscles contract to help stabilize the spine and keep it in a neutral position. This type of muscle action, where the muscle contracts and remains the same length, is called an isometric contraction. Answer: <answer>(C)</answer>.
Expert Comment: "Not only is this a good question the model got correct, its rationale and understanding of the muscle group and contraction type are also clear."

Supplementary Table 8 | Examples of multiple-choice questions incorrectly answered by PH-LLM. Separate examples for sleep and fitness are shown. In green, we provide a comment from the sleep and fitness expert about the model response.

Examples of multiple-choice questions incorrectly answered by PH-LLM

Sleep Question:
A patient undergoes polysomnography. During quiet wakefulness, no posterior alpha rhythm is observed when the eyes are closed. Which of the following statements is correct?
(A) With no clear posterior dominant alpha rhythm, sleep stage N1 cannot be definitively determined.
(B) If average background frequency during wakefulness is 8Hz, stage N1 can be defined only if activity decreases to $\leq$ 6Hz.
(C) Stage N1 can be scored if EEG demonstrates sharply contoured waves that are most prominent over central region and last <0.5 seconds.
(D) Sleep stage N1 can be scored with rapid eye movement.
(E) Sleep stage N1 can be scored with a brief series of waves at 12-14Hz, lasting 2 seconds and most prominent in central derivation.
Model Response:
The absence of alpha rhythm during quiet wakefulness suggests that the patient does not have a clear posterior dominant alpha rhythm. This makes it difficult to definitively determine sleep stage N1, as the presence of alpha rhythm is one of the key criteria for scoring N1. Answer: <answer>(A)</answer>
Correct Answer: (C)
Expert Comment: "This incorrect answer highlights 2 primary issues: 1. A lack of appropriate factual knowledge and/or application of the alternate criteria outlined in The AASM Manual for the Scoring of Sleep and Associated Events and 2. A lack of "common sense" reasoning that recognizes that the $\sim$10% of people without a posterior dominant rhythm also have N1 sleep (implying there must be other ways of scoring it)."

Fitness Question:
What shoulder movements and planes of action are involved in a tennis backhand?
(A) Adduction/frontal
(B) Abduction/frontal
(C) Adduction/transverse
(D) Abduction/transverse
Model Response: The tennis backhand involves horizontal adduction, which is movement toward the body in the transverse plane. Answer: <answer>(C)</answer>
Correct Answer: (D)
Expert Comment: "This indicates that the model is working out the mechanics. It gets the planes of movement, and terms of movement, but missed the 'backhand movement' that changes answer to abduction. It simply missed the most human element of the question."

Supplementary Table 9 | Case study evaluation rubric. 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

Supplementary Table 10 | PH-LLM example response for sleep case study on the insights section. Prompt in Supplementary Table 42 was used with `<sleep logs table>` shown in Supplementary Table 26, `<sleep summary>` shown in Supplementary Table 27, and `<gender> = Male, <age> = [20-25]` .

PH-LLM example response for sleep case study on the insights section.

Alertness

- The user's sleep logs show evidence of 4 naps, with an average nap length of 129.25 minutes. This suggests that the user may be experiencing issues with maintaining daytime alertness, or that they have an increased need for sleep beyond the usual 7-8 hours per night.

Routine and Schedule

- The user's bedtimes vary significantly, ranging from 16:10 to 06:22. Their wake times also vary significantly, ranging from 19:28 to 17:21. This inconsistent sleep-wake schedule can disrupt the body's natural circadian rhythm and lead to difficulties with sleep quality and daytime alertness.
- The midsleep point standard deviation is 04:26, which is high. This indicates that the user's sleep timing varies significantly from day to day. This can make it difficult to maintain a consistent sleep-wake schedule.

Sleep Duration

- The user's average sleep duration is 6 hours and 9 minutes, which is below the recommended range of 7-9 hours for adults. This suggests that the user may not be getting enough sleep, which can lead to a range of health problems, including fatigue, difficulty concentrating, irritability, and increased risk of accidents.
- The user's average sleep duration varies significantly from day to day, as noted by the sleep duration standard deviation of 2 hours and 39 minutes. This inconsistent sleep duration can further disrupt the body's natural circadian rhythm and lead to difficulties with sleep quality and daytime alertness.

Timing of the Main Sleep Period

- The user's average midsleep point (chronotype) is 04:00, which is within the normal range. This indicates that the user is neither an early bird nor a night owl.
- The user's average bedtime is 00:26 and their average wake time is 07:34. These are both slightly later than typical compared to similar users.

Sleep Efficiency

- The user's average sleep efficiency is 86%, which is slightly below the recommended range of 85-90%. This suggests that the user is spending a small but significant amount of time in bed awake.
- The user's average time to quality sleep is 33 minutes, which is longer than typical when compared to similar users. This may indicate that the user is having trouble falling asleep.
- The user's average WASO, or wake after sleep onset, is 10.08 minutes, which is typical when compared to similar users. This suggests that the user is not experiencing significant problems with staying asleep once they fall asleep.

Sleep Quality

- The user's average light sleep duration is 3 hours and 50 minutes, which is less than typical when compared to similar users.
- The user's average deep sleep duration is 1 hour and 8 minutes, which is typical when compared to similar users.
- The user's average REM sleep duration is 1 hour and 10 minutes, which is typical when compared to similar users.
- The user's average restlessness is 13%, which is higher than typical when compared to similar users. This may indicate that the user is experiencing difficulties with sleep quality, such as frequent arousals or awakenings.

Sleep Architecture

- The user's sleep architecture appears to be relatively normal, with an appropriate distribution of light, deep, and REM sleep.

Other insights

- The user's heart rate during sleep ranges from 58 to 63 bpm, which is within the normal range for adults.

Supplementary Table 11 | PH-LLM example response for sleep case study on the etiology section. Prompt in Supplementary Table 43 and model answer from previous insight section was used as input. For `<insights response>`, we use the response in Supplementary Table 10, for `<sleep summary>` we again use Supplementary Table 27, and `<gender> = Male, <age> = [20-25]`.

PH-LLM example response for sleep case study on the etiology section.

Based on your data, there are some areas for improving your sleep that we can explore.

Circadian Rhythm:

- Your sleep schedule is inconsistent, which can make it difficult for your body to develop a consistent rhythm. Your average bedtime ranges from 16:10 to 06:22 and your average wake time ranges from 19:28 to 17:21. Your midsleep point standard deviation is 4 hours and 26 minutes which is very high. This indicates that your sleep timing is highly variable. This makes it difficult for your body to develop a consistent pattern.

Homeostatic Drive:

- Your average sleep duration is 6 hours and 9 minutes which is less than the recommended 7-8 hours for adults. You have 4 naps during the recorded period for an average of 129.25 minutes. This could be a sign that you have insufficient homeostatic drive.

Psychological Hyperarousal:

- Your restlessness metric is 13% which is higher than normal when compared to similar users. This could be a sign that you have psychophysiologic hyperarousal, a condition characterized by an overactive mind and body that can interfere with sleep.

Extrinsic Factors:

- There is insufficient documentation to comment on this.

Supplementary Table 12 | PH-LLM example response for sleep case study on the recommendations section. Prompt in Supplementary Table 44 and model answers from previous insight and etiology sections were used as input. For `<insights response>`, we use the response in Supplementary Table 10, for `<etiology response>`, we use the response in Supplementary Table 11, for `<sleep summary>` we again use Supplementary Table 27, and `<gender> = Male, <age> = [20-25]`.

PH-LLM example response for sleep case study on the recommendations section.

Based on your data, there are some areas for improving your sleep:

Sleep schedule:

- Your sleep schedule is inconsistent, which can make it difficult for your body to develop a consistent rhythm.
- Your average midsleep point standard deviation is 4 hours and 26 minutes which is very high. This indicates that your sleep timing is highly variable. This makes it difficult for your body to develop a consistent pattern.
- A goal sleep schedule of 11:30 PM to 7:15 AM would be helpful for your current work schedule and provide you with approximately 7 hours 45 minutes of sleep each night.
- Gradually adjust your bedtime and wake time by 15-30 minutes every few days until you reach your goal.
- Be patient and persistent, as it may take some time to adjust to your new sleep schedule.
- Remember, making small changes over time is more effective than trying to change everything at once.
- Try to reduce napping, and instead add this time to your primary sleep period.
- If you have any trouble sticking to your sleep goal, don't be discouraged. Just keep trying and you will eventually succeed.

Sleep Duration:

- Your average sleep duration is 6 hours and 9 minutes which is less than the recommended 7-8 hours for adults.
- You have 4 naps during the recorded period for an average of 129.25 minutes. This could be a sign that you have insufficient homeostatic drive.
- Make sure you're getting enough physical activity during the day, as regular exercise can help you fall asleep more easily at night.
- Avoid working out too close to bedtime, as this can be stimulating and make it harder to fall asleep.
- Be sure to get sunlight exposure during the day, as this can help regulate your body's natural sleep-wake cycle.
- Avoid caffeine and alcohol in the hours leading up to bed, as these substances can interfere with sleep.
- If you find yourself waking up in the middle of the night, don't be discouraged. This is normal. Just relax and try to clear your head. If you can't fall back asleep after 20 minutes, get out of bed and do something calming until you feel tired.

Sleep Quality:

- Your restlessness metric is 13% which is higher than normal when compared to similar users. This could be a sign that you have psychophysiologic hyperarousal, a condition characterized by an overactive mind and body that can interfere with sleep.
- Create a sleep-wake schedule that is consistent on both weekdays and weekends, and stick to it as closely as possible. This will help to regulate your circadian rhythm and improve your sleep quality.
- Set aside at least 1 hour before bed to help your body wind down in a dimly lit, low-stimulation setting can help prepare you for that early bedtime.
- Consider relaxation techniques, such as deep breathing exercises, yoga, or meditation, before bed to reduce arousal.
- Make sure your bedroom environment is conducive to sleep, meaning it is dark, quiet, and cool.

Followup Question

- Do you know how long it takes to lose or gain a day after shifting your sleep time to later or earlier.

Supplementary Table 13 | PH-LLM example response for fitness case study on the demographics section. Prompt in Supplementary Table 45 was used with `<age> = [40-45]`, `<height> = 1.80m`, `<weight> = 84.53kg`, `<BMI> = 26.0`, and `<gender> = Male`.

PH-LLM example response for fitness case study on the demographics section.

Insights:

- The trainee is a [40-45]-year-old male with a BMI of 26.0, indicating that he is slightly overweight.
- No significant health concerns or exceptional demographics are apparent from the provided data.

Special Precautions and Considerations:

- Gradual Progression: It is important to start slowly and gradually increase the intensity and duration of workouts over time to avoid injury.
- Listen to the Body: Pay attention to any pain or discomfort during exercise and stop if necessary.
- Hydration: Staying hydrated before, during, and after workouts is crucial to prevent dehydration.

Supplementary Table 14 | PH-LLM example response for fitness case study on the training load section. Prompt in Supplementary Table 46 was used with `<table of daily activity metrics>` shown in Supplementary Table 28, `<aggregate statistics of daily activity metrics>` shown in Supplementary Table 29, and `<day of the week> = Thursday`. We show specific dates only as `the Nth day` for privacy.

PH-LLM example response for fitness case study on the training load section.

Training Load Trends and Patterns

Observation: The trainee's average daily moderate activity (Fat-burn zone) over the last 30 days is 12.3 minutes, while vigorous activity (Cardio and Peak zones) averages 12.7 minutes.
Insight: The trainee engages in a balanced mix of moderate and vigorous activities.
Observation: The acute TRIMP (7-day total training load) is 346, and the chronic TRIMP (28-day average acute training load) is 235.
Insight: The trainee's training load has been relatively consistent over the past month.
Observation: The Acute-Chronic Workload Ratio (ACWR) is 1.5.
Insight: The trainee is currently pushing their training load limits, which may increase the risk of injury or overtraining.

Intensity and Duration

Observation: The trainee has spent a significant amount of time in the Fat-burn zone (50% heart rate reserve) during their workouts, indicating a focus on moderate-intensity activities.
Insight: This is beneficial for building cardiovascular fitness and burning fat.
Observation: There have been fewer instances of time spent in the Cardio zone (70% heart rate reserve) and Peak zone (85% heart rate reserve) recently.
Insight: The trainee may be focusing on lower-intensity activities or incorporating more rest periods into their workouts.
Observation: The trainee's most significant workout was a 46-minute treadmill session on the 30th day, where they spent 14 minutes in the Peak zone.
Insight: This intense workout contributed significantly to the trainee's overall training load for the week.

Frequency and Rest Periods

Observation: The trainee has been working out consistently, with an average of 5-6 workouts per week.
Insight: This frequency is sufficient for maintaining fitness and improving cardiovascular health.
Observation: There have been a few rest days scattered throughout the past month.
Insight: Incorporating rest days is crucial for recovery and preventing burnout.

Recent Activity

Observation: Today (the 30th day), the trainee has already completed 5 workouts, including a 46-minute treadmill session, a 15-minute walk, and a 10-minute walk.
Insight: The trainee is having an exceptionally active day, which may contribute to their elevated acute training load.

Conclusion

Overall, the trainee has been engaging in a balanced mix of moderate and vigorous activities. Their training load has been relatively consistent over the past month, and they are currently pushing their limits. It is important for the trainee to listen to their body and rest when needed to avoid overtraining.

Supplementary Table 15 | PH-LLM example response for fitness case study on the sleep section. Prompt in Supplementary Table 47 was used with `<table of sleep metrics for fitness case studies>` shown in Supplementary Table 30, `<aggregate statistics of sleep metrics>` shown in Supplementary Table 31, and `<day of the week> = Thursday`.

PH-LLM example response for fitness case study on the sleep section.

Sleep Schedule

Observations:

- The trainee's sleep schedule is highly inconsistent, with bedtimes ranging from 22:13 to 02:16 and wake times ranging from 03:26 to 11:30.
- The trainee's average bedtime is 00:11 and their average wake time is 07:35.

Insights:

- The inconsistent sleep schedule is likely to be disruptive to the trainee's circadian rhythm, which can lead to sleep problems and other health issues.

Sleep Duration

Observations:

- The trainee's average sleep duration over the last 30 days is 6.3 hours.
- The trainee's sleep duration has varied significantly from night to night, ranging from 2.95 hours to 8.08 hours.
- The trainee's sleep duration has been below the recommended 7-9 hours for adults on 22 out of the last 30 days.

Insights:

- The trainee's short sleep duration is a cause for concern, as it can lead to a number of health problems, including obesity, heart disease, diabetes, and depression.
- The trainee's inconsistent sleep duration is likely to be contributing to their poor sleep quality.

Sleep Quality

Observations:

- The trainee's average sleep score over the last 30 days is 76.
- The trainee's sleep score has varied significantly from night to night, ranging from 48.54 to 87.57.
- The trainee's sleep score has been below the optimal 85 on 24 out of the last 30 days.

Insights:

- The trainee's sleep quality is generally poor.
- The trainee's low sleep scores may be due to a number of factors, including their short sleep duration, inconsistent sleep schedule, and poor sleep hygiene.

Recent Trends

Observations:

- The trainee's sleep duration has been slightly below their monthly average in the last 3 days.
- The trainee's sleep quality has been slightly below their monthly average in the last 3 days.

Insights:

- The trainee's recent trends suggest that they are not getting enough sleep and their sleep quality is poor.

Today's Sleep

Observations:

- The trainee went to bed at 23:57 and woke up at 03:26 today.
- The trainee's sleep duration today was 2.95 hours, which is significantly lower than their monthly average.
- The trainee's sleep score today was 62.9, which is significantly lower than their monthly average.

Insights:

- The trainee's sleep last night was very short and of poor quality.
- The trainee is likely to be tired and groggy today.

Supplementary Table 16 | PH-LLM example response for fitness case study on the health metrics section. Prompt in Supplementary Table 48 was used with `<table of health metrics over the past 30 days>` shown in Supplementary Table 32, `<table of health metrics over the past week>` shown in Supplementary Table 33, `<aggregate statistics of health metrics>` shown in Supplementary Table 34, and `<day of the week> = Thursday`.

PH-LLM example response for fitness case study on the health metrics section.

Resting Heart Rate

- Baseline: The trainee's average resting heart rate over the past month is 59.03 bpm.
- Recent trend: Resting heart rate has increased in the past week, from 54 bpm to 66 bpm.
- Today's value: 66 bpm, which is higher than the baseline and the recent trend.

Insight: The elevated resting heart rate today could indicate increased stress, poor sleep, or insufficient recovery.

Heart Rate Variability (HRV RMSSD)

- Baseline: The trainee's average HRV RMSSD over the past month is 24.11 ms.
- Recent trend: HRV RMSSD has fluctuated in the past week, ranging from 15.95 ms to 28.23 ms.
- Today's value: 18.17 ms, which is lower than the baseline and the recent average.

Insight: The low HRV RMSSD today suggests reduced parasympathetic activity and may indicate a state of stress or insufficient recovery.

Respiratory Rate

- Baseline: The trainee's average respiratory rate over the past month is 15.18 breaths/minute.
- Recent trend: Respiratory rate has remained relatively stable in the past week, ranging from 15.40 to 16.80 breaths/minute.
- Today's value: 16.00 breaths/minute, which is consistent with the baseline and recent trend.

Insight: The stable respiratory rate suggests no significant changes in respiratory function or overall health status.

Today's Assessment

- The elevated resting heart rate and low HRV RMSSD today indicate a state of stress or insufficient recovery.
- It is important for the trainee to prioritize rest and recovery activities, such as getting enough sleep, managing stress, and allowing sufficient time for recovery between workouts.
- Monitoring these metrics regularly can help the trainee track their recovery and make adjustments to their training and lifestyle as needed.

Supplementary Table 17 | PH-LLM example response for fitness case study on the assessment section. Prompt in Supplementary Table 49 and model answers for previous sections were used as input. For `<demographics response>`, we use the response in Supplementary Table 13, for `<training load response>` we use Supplementary Table 14, for `<sleep metrics response>`, we use Supplementary Table 15, for `<health metrics response>`, we use Supplementary Table 16. We also incorporate synthetically generated qualitative feedback from the user with `<subjective readiness>` shown in Supplementary Table 35 and `<muscle soreness>` shown in Supplementary Table 36.

PH-LLM example response for fitness case study on the assessment section.

Important insights
Load

- The trainee's acute training load is 346 and their ACWR is 1.5. This is consistent with the last month. This indicates a slightly high ACWR which may put the trainee at risk for over training.

Sleep

- The trainee's average sleep duration is 6.3 hours with a sleep score of 76. This is below the recommended 7-9 hours. The trainee has been below the recommended amount of sleep for 22 of the last 30 days. Today's sleep was 2.95 hours and a score of 62.9 which is significantly lower than the monthly averages.
- The trainee's sleep schedule is inconsistent.

Health Metrics

- The trainee's resting heart rate is 66 bpm which is elevated compared to the monthly average of 59.03 bpm. This could indicate stress, poor sleep, or insufficient recovery.
- The trainee's HRV is 18.17 ms which is decreased compared to the monthly average of 24.11 ms. This could indicate stress, poor sleep, or insufficient recovery.

Subjective Readiness + Muscle Soreness

- The trainee is feeling a bit stressed and fatigued but is managing by hydrating and prioritizing recovery.
- The trainee is feeling some soreness in their legs but it is manageable.

Readiness Score

- 2/5
- Explanation: The trainee's load and metrics indicate a need for recovery. The trainee's sleep has been poor and inconsistent as well. The trainee's subjective assessment also suggests a need for recovery.

Fitness Recommendations for Today

- The trainee should prioritize recovery today by taking a rest day or doing a light activity.
- The trainee should focus on getting a good night's sleep by going to bed and waking up at the same time each day, creating a relaxing bedtime routine, and making sure their bedroom is dark, quiet, and cool.
- The trainee should manage stress by incorporating stress-reducing activities into their day, such as exercise, yoga, or meditation.
- The trainee should ensure they are hydrating adequately by drinking plenty of water throughout the day.

Followup Questions

- What happened to sleep last night?
- What caused the trainee to feel more stressed and fatigued?

Supplementary Table 18 | PH-LLM example response for fitness case study on the health metrics section with 5 days of data. The original data shown in Supplementary Table 28 was used, but only including the last 5 days. Based on this reduced data aggregate statistics were computed similar to Supplementary Table 29. In green, we highlight parts of the response that remained similar, and in blue, we highlight parts of the response that changed between all versus just 5 days of data. The original PH-LLM response with 30 days is in Supplementary Table 16. Some conclusions remain consistent across both timeframes. For example, the rising trend in resting heart rate is evident in both 5-day and 30-day analyses, leading to similar recommendations about potential stress or insufficient recovery. However, other conclusions change significantly based on data availability. With 5 days of data, today's heart rate variability (HRV) of 18.17 ms appears slightly above the period average (17.41 ms), suggesting improved recovery. In contrast, the 30-day analysis reveals this value is well below the monthly average (24.11 ms), identifying a state of increased stress (Supplementary Table 16).

PH-LLM example response for fitness case study on the health metrics section with only 5 days of data.

Resting Heart Rate (RHR)

Observations:

- The trainee's average RHR for the last 5 days is 64 bpm.
- The RHR has been steadily increasing since Sunday, from 60 bpm to 66 bpm.
- Today's RHR of 66 bpm is higher than the average for the last 5 days.

Insights:

- The increase in RHR over the last few days may indicate increased stress or insufficient recovery.
- Today's RHR is slightly elevated, which may be due to factors such as dehydration or lack of sleep.

Heart Rate Variability (HRV)

Observations:

- The trainee's average HRV RMSSD for the last 5 days is 17.41 ms.
- HRV has been fluctuating over the last 5 days, ranging from 15.95 ms to 19.04 ms.
- Today's HRV of 18.17 ms is higher than the average for the last 5 days.

Insights:

- The fluctuations in HRV may indicate varying levels of stress and recovery.
- Today's HRV is slightly above the average, which may indicate improved recovery or decreased stress.

Respiratory Rate (RR)

Observations:

- The trainee's average RR for the last 5 days is 16.04 breaths/minute.
- RR has remained relatively stable over the last 5 days, ranging from 15.40 breaths/minute to 16.80 breaths/minute.
- Today's RR of 16.00 breaths/minute is consistent with the average for the last 5 days.

Insights:

- The stable RR indicates consistent breathing patterns.
- Today's RR is within the normal range and does not indicate any significant changes in respiratory function.

Supplementary Table 19 | PH-LLM example response on the training load section with modified statistics. We change the following statistics in Supplementary Table 29: Acute TRIMP (7-day total training load): 235, Chronic TRIMP (28-day average acute training load): 346, Acute-Chronic Workload Ratio (ACWR): 0.6. Model response portions relevant to the change are highlighted in blue.

PH-LLM example response on the training load section with modified statistics.

Training Load Trends and Patterns

Observation: The trainee's average daily moderate activity (Fat-burn zone) over the last 30 days is 12.33 minutes, while their average vigorous activity (Cardio and Peak zones) is 12.67 minutes.
Insight: This indicates a balanced distribution between moderate and vigorous activity levels, which is generally recommended for overall fitness.
Observation: The trainee's acute TRIMP (7-day total training load) is 235, while their chronic TRIMP (28-day average acute training load) is 346.
Insight: This suggests that the trainee's training load has decreased recently compared to their average load over the past month.
Observation: The trainee's Acute-Chronic Workload Ratio (ACWR) is 0.6.
Insight: This ratio is well below the recommended threshold of 1.5, indicating that the trainee is at a low risk of injury due to overtraining.

Intensity

Observation: The trainee has been spending a significant amount of time in the Fat-burn zone during their workouts, with an average of 12.33 minutes per day.
Insight: This suggests that the trainee is focusing on moderate-intensity activities, which are beneficial for improving cardiovascular health and endurance.
Observation: The trainee has also been spending time in the Cardio and Peak zones, with an average of 6.64 minutes and 6.03 minutes per day, respectively.
Insight: These higher-intensity activities are important for building strength, power, and improving overall fitness.

Duration

Observation: The trainee's workouts have been relatively consistent in duration over the last 30 days, with an average of 19.2 minutes.
Insight: This suggests that the trainee is able to maintain a good level of activity for a sustained period of time.
Observation: The trainee's longest workout in the last 30 days was a 46-minute treadmill session.
Insight: This workout likely contributed significantly to the trainee's training load and overall fitness.

Frequency

Observation: The trainee has been working out consistently throughout the week, with no significant gaps in activity.
Insight: This consistency is important for maintaining fitness and making progress towards training goals.

Rest Periods

Observation: The trainee's activity metrics and exercise logs show that they have been taking regular rest days.
Insight: These rest days are essential for recovery and preventing burnout.

Recent Activity Levels

Observation: The trainee's activity levels have been relatively high in the last few days, with a significant increase in TRIMP.
Insight: This may be due to several factors, such as increased training intensity or duration.

Significant Workouts

Observation: The trainee completed a 46-minute treadmill session on Thursday, the 30th day.
Insight: This workout was significantly longer and more intense than their other recent workouts and likely contributed to the increase in their training load.
Observation: The trainee also completed several shorter walks on the same day, totaling an additional 90 minutes of activity.
Insight: These walks likely contributed to the trainee's overall time in the Fat-burn zone and helped them recover from the treadmill session.

Supplementary Table 20 | PH-LLM example response on the sleep section with missing sleep metrics for today. We replace the data for today in Supplementary Table 30 with "Thursday the 30th day NaN NaN NaN NaN NaN NaN NaN". Model response portions relevant to the change are highlighted in blue.

PH-LLM example response on the sleep section with missing sleep metrics for today.

Sleep Schedule and Duration

- The trainee's sleep schedule is relatively consistent, with an average bedtime of 00:11 and an average wake-time of 07:35.
- The trainee's average sleep duration over the last 30 days is 6.3 hours, which is significantly less than the recommended 7-9 hours per night for adults.
- The trainee's sleep duration varies significantly from day to day, with a standard deviation of 1.3 hours.
- In the last 3 days, the trainee's sleep duration has been significantly lower than their monthly average, with a Z-score of -0.6.
- In the last 2 days, the trainee's sleep duration has been significantly lower than their monthly average, with a Z-score of -1.6.
- The trainee's sleep duration has been variable over the past 30 days, with a standard deviation of 1.3 hours.
- The trainee's sleep duration has been particularly low in the last few days, with a Z-score of -0.6.
- The trainee's sleep duration has been consistently below the recommended 7-9 hours per night for adults, with a mean of 6.3 hours.

Insights:

- The trainee's short and variable sleep duration may be negatively impacting their physical and mental health.
- The trainee should consider making changes to their sleep habits to improve the quantity and quality of their sleep.

Sleep Quality

- The trainee's average sleep score over the last 30 days is 76, which is considered to be good.
- The trainee's sleep score varies significantly from day to day, with a standard deviation of 9.1.
- In the last 3 days, the trainee's sleep score has been significantly lower than their monthly average, with a Z-score of -0.2.

Insights:

- The trainee's variable sleep quality may be due to a number of factors, such as stress, anxiety, or poor sleep hygiene.
- The trainee should consider making changes to their sleep habits to improve the quality of their sleep.

Today's Sleep

- Today, the trainee's sleep duration is missing.
- Today, the trainee's sleep score is missing.

Insights:

- It is not possible to comment on the trainee's sleep today without data.

Supplementary Table 21 | AutoEval model performance on 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 (low variance)	0.284 (0.230–0.336)	0.278 (0.225–0.329)	0.142 (0.106–0.182)
Fitness Primary C (high variance)	0.315 (0.269–0.360)	0.301 (0.256–0.343)	0.321 (0.274–0.366)
Sleep			
Gemini Pro	0.242 (0.185–0.299)	0.230 (0.177–0.283)	0.223 (0.166–0.280)
Sleep Primary D (low variance)	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 (high variance)	0.355 (0.307–0.401)	0.324 (0.280–0.367)	0.367 (0.316–0.419)

Supplementary Table 22 | 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)

Supplementary Table 23 | 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 was statistically significant. The difference in AUROC, the corresponding 95% confidence intervals, and the statistical significance were calculated via a paired bootstrap estimator across 100 bootstrapping iterations.

	PH-LLM Zero-shot	PH-LLM Few-shot	Logistic Regression
Very restless	**0.162** (0.084–0.253)	0.133 (-0.009–0.236)	0.043 (-0.014–0.093)
Satisfied	**0.138** (0.042–0.225)	**0.168** (0.085–0.240)	-0.018 (-0.068–0.040)
Refreshed	**0.113** (0.030–0.203)	**0.120** (0.050–0.205)	0.012 (-0.048–0.069)
Trouble falling asleep	**0.219** (0.088–0.324)	**0.161** (0.054–0.264)	0.033 (-0.030–0.095)
Trouble staying asleep	0.103 (-0.001–0.204)	0.075 (-0.016–0.164)	-0.037 (-0.112–0.034)
Trouble sleeping	**0.158** (0.053–0.240)	0.117 (-0.003–0.213)	-0.021 (-0.093–0.063)
Enough sleep	**0.152** (0.027–0.284)	**0.233** (0.079–0.359)	-0.027 (-0.123–0.078)
Quality	0.112 (-0.070–0.260)	**0.245** (0.085–0.419)	0.085 (-0.041–0.195)
Trouble being productive	0.020 (-0.142–0.188)	**0.182** (0.029–0.331)	0.014 (-0.060–0.098)
Alert	**0.093** (0.019–0.150)	**0.093** (0.026–0.161)	0.004 (-0.038–0.053)
Tiredness	**0.163** (0.107–0.229)	**0.145** (0.075–0.206)	0.017 (-0.016–0.050)
Having problems	**0.198** (0.046–0.305)	0.011 (-0.105–0.139)	-0.020 (-0.094–0.053)
SI due to trouble concentrating	**0.253** (0.133–0.367)	**0.158** (0.022–0.258)	-0.024 (-0.073–0.025)
SI due to irritability	0.119 (-0.054–0.337)	**0.186** (0.031–0.373)	-0.092 (-0.167–0.005)
Sleepy during daytime	**0.144** (0.066–0.237)	**0.108** (0.019–0.186)	0.005 (-0.028–0.039)
Trouble staying awake	**0.158** (0.047–0.298)	0.086 (-0.085–0.239)	0.019 (-0.073–0.094)

Supplementary Table 24 | 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 was statistically significant. The difference in AUPRC, the corresponding 95% confidence intervals, and the statistical significance were calculated via a paired bootstrap estimator across 100 bootstrapping iterations.

	PH-LLM Zero-shot	PH-LLM Few-shot	Logistic Regression
Very restless	**0.067** (0.029–0.122)	0.042 (-0.010–0.113)	0.006 (-0.041–0.068)
Satisfied	**0.065** (0.009–0.135)	**0.059** (0.003–0.132)	-0.046 (-0.094–0.021)
Refreshed	0.031 (-0.009–0.073)	0.031 (-0.005–0.073)	-0.022 (-0.062–0.015)
Trouble falling asleep	**0.065** (0.024–0.118)	**0.052** (0.004–0.099)	-0.031 (-0.087–0.013)
Trouble staying asleep	0.016 (-0.038–0.074)	0.024 (-0.012–0.074)	-0.067 (-0.141–0.009)
Trouble sleeping	**0.036** (0.010–0.078)	0.020 (-0.023–0.061)	**-0.072** (-0.168—0.012)
Enough sleep	0.044 (-0.004–0.105)	**0.058** (0.014–0.118)	-0.035 (-0.122–0.014)
Quality	0.013 (-0.013–0.047)	**0.025** (0.004–0.056)	0.004 (-0.020–0.025)
Trouble being productive	-0.014 (-0.070–0.022)	0.009 (-0.055–0.060)	0.007 (-0.017–0.040)
Alert	**0.063** (0.022–0.098)	0.051 (-0.007–0.092)	0.002 (-0.031–0.039)
Tiredness	**0.063** (0.020–0.106)	**0.066** (0.033–0.101)	0.023 (-0.009–0.049)
Having problems	0.010 (-0.044–0.040)	-0.023 (-0.107–0.018)	-0.005 (-0.026–0.015)
SI due to trouble concentrating	**0.048** (0.015–0.087)	0.035 (-0.003–0.079)	-0.051 (-0.138–0.009)
SI due to irritability	0.021 (-0.019–0.085)	0.038 (-0.001–0.092)	-0.031 (-0.088–0.018)
Sleepy during daytime	**0.036** (0.001–0.081)	0.026 (-0.025–0.083)	0.001 (-0.028–0.023)
Trouble staying awake	**0.017** (0.002–0.045)	-0.000 (-0.044–0.035)	-0.008 (-0.043–0.009)

Supplementary Table 25 | Comparison of demographic distributions. United States Census statistics were retrieved from `https://data.census.gov`. Distributions for Ethnicity and Race were drawn from the 2020: DEC Demographic Profile, Biological Sex and Age were drawn from the 2023 American Community Survey 1-Year estimates, and Orientation was drawn from 2021 Household Pulse Surveys. Fitbit data distributions were drawn from sleep and fitness case studies we collected, data from the study on patient-reported outcomes, and anonymized Fitbit production data from individuals who provided consent for research purposes ("Fitbit Pop."). Empty cells indicate the option was not measured in that population.

Variable	**Definition**	United States Census	PH-LLM Case Studies (Sleep)	PH-LLM Case Studies (Fitness)	PH-LLM Patient-Reported Outcomes	Fitbit Pop.
Ethnicity	Hispanic	19%			8%	
	Not hispanic	81%			92%	
Race*	White	69%			84%	
	Black	14%			5%	
	Asian	7%			3%	
	Native American	1%			1%	
	Another race	9%			7%	
Biological Sex	Male	49%	49%	72%	25%	46%
	Female	51%	51%	28%	72%	54%
	I Prefer Something Else				3%	
Age†	<30	24%‡	17%	3%	10%	18%
	30 - 39	17%	14%	22%	27%	23%
	40 - 49	16%	17%	36%	27%	20%
	50 - 59	14%	14%	17%	20%	16%
	60 - 69	14%	16%	16%	12%	13%
	70 - 79	10%	16%	3%	4%	8%
	80+	5%	6%	2%	<1%	3%
Orientation	Heterosexual	88%			77%	
	LGBTQIA+	8%			19%	
	I Prefer Something Else	4%			4%	

$*$ US Census percentages were based on all individuals reporting a single race, with "Native Hawaiian and Other Pacific Islander alone" and "Some Other Race Alone" combined into the "Another race" category.
† US Census percentages were based on all individuals aged 15 or older for comparability to the reported Fitbit studies.
‡ This bucket spans the ages $15 - 29$ owing to available granularity of the US Census data.

Supplementary Table 26 | 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
1st 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
2nd 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
3rd 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
4th 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
5th 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
5th 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
6th 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
7th 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
7th 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

Supplementary Table 27 | 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 versus 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

Supplementary Table 28 | 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	1st day	15.0	27.0	0.0	62.0	16200
Thursday	2nd day	19.0	23.0	1.0	62.0	9900
Friday	3rd day	6.0	0.0	0.0	6.0	5950
Saturday	4th day	20.0	0.0	0.0	20.0	11210
Sunday	5th day	1.0	0.0	0.0	1.0	8160
Monday	6th day	7.0	0.0	0.0	7.0	13120
Tuesday	7th day	12.0	0.0	0.0	12.0	15490

Supplementary Table 29 | 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 the 29th 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 the 29th 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 the 30th 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

Supplementary Table 30 | 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	1st day	23:01	07:05	7	53	80	18	80
Thursday	2nd day	22:48	07:17	7	49	94	17	84
Friday	3rd day	22:43	07:12	7	77	61	13	71
Saturday	4th day	00:15	08:12	7	55	87	21	83
Sunday	5th day	01:11	09:33	7	62	86	15	74
Monday	6th day	23:16	07:31	7	57	104	19	86
Tuesday	7th day	22:13	04:04	4	55	41	13	64

Supplementary Table 31 | 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

Supplementary Table 32 | 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	1st day	53.0	27	14
Thursday	2nd day	54.0	22	13
Friday	3rd day	55.0	27	13
Saturday	4th day	56.0	23	15
Sunday	5th day	57.0	23	14
Monday	6th day	56.0	31	14
Tuesday	7th day	56.0	19	15
Wednesday	8th day	58.0	NaN	NaN
Thursday	9th day	61.0	17	15
Friday	10th day	64.0	13	15
Saturday	11th day	62.0	23	15
Sunday	12th day	63.0	16	15
Monday	13th day	62.0	26	14
Tuesday	14th day	60.0	28	14
Wednesday	15th day	61.0	17	15
Thursday	16th day	59.0	30	14
Friday	17th day	57.0	35	15
Saturday	18th day	58.0	25	16
Sunday	19th day	58.0	20	16
Monday	20th day	60.0	16	15
Tuesday	21st day	58.0	29	14
Wednesday	22nd day	56.0	40	13
Thursday	23rd day	54.0	41	14
Friday	24th day	56.0	28	15
Saturday	25th day	57.0	NaN	NaN
Sunday	26th day	60.0	17	16
Monday	27th day	62.0	15	15
Tuesday	28th day	65.0	19	16
Wednesday	29th day	67.0	16	16
Thursday	30th day	66.0	18	16

Supplementary Table 33 | Example of health metrics table over the past week for a particular individual used in fitness case studies.

Example of health metrics table over the past week 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	24th day	56.0	28	15
Saturday	25th day	57.0	NaN	NaN
Sunday	26th day	60.0	17	16
Monday	27th day	62.0	15	15
Tuesday	28th day	65.0	19	16
Wednesday	29th day	67.0	16	16
Thursday	30th day	66.0	18	16

Supplementary Table 34 | 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
Standard deviation Resting Heart Rate: 3 bpm
Resting Heart Rate Z-score: 1.9
Mean HRV RMSSD: 24 ms
Standard deviation HRV RMSSD: 7 ms
HRV RMSSD Z-score: -0.8
Mean Respiratory Rate: 15 breaths/minute
Standard deviation Respiratory Rate: 0.83 breaths/minute
Respiratory Rate Z-score: 0.9
Past week:
Resting Heart Rate range: 56 to 67 bpm
HRV RMSSD range: 16 to 28 ms
Respiratory Rate range: 15 to 17 breaths/min

Supplementary Table 35 | 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.

Supplementary Table 36 | 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.

Supplementary Table 37 | Prompt for generating synthetic user input about subjective readiness for fitness case studies. Model answers from previous sections (training load, sleep metrics, and health metrics) were used as input to personalize the synthetic user feedback.

Prompt for generating synthetic user input about subjective readiness for fitness case studies.

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

<sleep metrics response>

<health metrics response>

Some example subjective readiness responses are:
3/5 - There's a lot going on, but I'm keeping my head above water.
4/5 - Things are busy, but I'm managing the stress well.
2/5 - I've got some family things that are weighing on my mind.
2/5 - Not fully energized; it's been a long week.
3/5 - Feeling somewhat fatigued but nothing out of the ordinary.
5/5 - I got plenty of rest, feeling zesty and spirited!
4/5 - Almost back to full strength after the last session.
1/5 - Really feeling the effects of our last intense workout.
3/5 - A bit of stiffness, but nothing that will stop me.
4/5 - Sleep was better than usual, but still not perfect.
1/5 - Hardly slept a wink, feeling pretty zombified.
3/5 - It took a while to fall asleep; a little groggy.
4/5 - Generally positive and feeling healthy.
2/5 - A bit under the weather, but I can push through.
5/5 - In top form mentally and physically; it's a great day.
2/5 - The soreness is noticeable; I'm aware of it with every movement.
4/5 - There's some residual tightness, but it's not slowing me down.
1/5 - My quads are really aching; I might need some extra recovery time.
5/5 - I've eaten well and stayed hydrated, feeling prepped.
3/5 - Meals were a bit off-schedule, but I've kept up with hydration.
2/5 - Not the best eating day, and I probably haven't drunk enough water.
5/5 - Sharp and ready, I can concentrate fully on the workout ahead.
3/5 - I'm a bit distracted with some personal issues but will try to focus on training.
2/5 - Having trouble concentrating, a lot on my mind today.
5/5 - Fired up and ready to give it my all today!
3/5 - I'm somewhat motivated; not at my peak but willing to work.
1/5 - Finding it hard to muster the enthusiasm, but I'm here.
2/5 - This heat is intense! I'm feeling pretty drained before starting.
5/5 - Perfect weather for me. I feel like I can run forever!
3/5 - It's a bit chillier than I like, but once I warm up, I should be fine.
3/5 - A bit slippery on the track, but still good to go.

Based on the above observations and insights determined based on a trainee's training load, sleep and health metrics data, write a hypothetical subjective readiness response by the trainee that is coherent with the above insights.

Consider statements about stress, fatigue, physical recovery, environmental/weather conditions, nutrition and hydration.

The response should be a single sentence, start with a number 1-5 that represents the Likert scale value of how ready the trainee is likely to feel.

Keep the response extremely short as if the trainee is responding in an app on a mobile phone.

Supplementary Table 38 | Prompt for generating synthetic user input about muscle soreness for fitness case studies. Model answers from previous sections (demographics, training load, and subjective readiness) were used as input to personalize the synthetic user feedback.

Prompt for generating synthetic user input about muscle soreness for fitness case studies

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

<training load response>

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

Some example muscle soreness responses are:
New to this, but no soreness yet, feeling good!
My legs are feeling it after starting jogging, but it's a good kind of pain.
Overdid it cycling, my thighs are really sore.
Usual tightness in my calves.
Legs are feeling strong, the usual mileage isn't phasing me.
Shoulders are a bit sore.
My upper back is aching.
Quads are holding up well, even after increasing my ride distance.
Feeling the burn in my calves.
Really pushed the pace and now my legs are paying the price, very sore.
Full-body workout has left me all-over sore, but it's manageable.
Quick recovery from last session, feeling agile and ready.
Did too many burpees; my arms and core are super sore.
Feeling loose and limber, no soreness after my yoga flow.
A little tightness in the hips, but it's nothing out of the ordinary.
Those new poses targeted muscles I don't usually work, feeling it today.
The usual DOMS, but I'll power through.
My grip and forearms are sore from all the lifting and pulling.
Overestimated my limits, my lower back is quite sore from deadlifts.
The long run took its toll on my knees; they're aching a bit today.
My feet and calves are really feeling the miles, need extra recovery time.
Arms feel ready, no soreness despite the heavy bag work.
Shoulders and back are sore, but it's nothing unusual.
Overdid the uppercuts, my arms are sore and my movement feels limited.
The pump is there, and so is the soreness, but it's the good kind.
Pushed too hard on the bench press; my chest and triceps are stiff.
Loving the new routine, no soreness, and feeling energized!
Shin splints are starting to flare up, but I'm learning to manage them.
Pushed for that extra mile; now my quads and calves are protesting.
Post-long run, my hamstrings are tight, but it's the usual fatigue.
A little stiffness in the ankles this morning, but it loosened up after a mile.
Really feeling yesterday's terrain change, especially in my Achilles tendon.
Surprisingly fresh after those sprints, ready to hit the track again.
My glutes are sore from explosive starts.
The intensity of the repeats got to my hamstrings, they're quite sore.
Some ankle soreness from the rocky paths, but I'm used to it.
Legs are strong, no soreness, even after that steep hill session.
Tripped up a bit yesterday; my knees and hips are feeling it today.
Feeling a slight fatigue in the legs but nothing major.
The mileage is starting to add up, my calves are asking for a day off.
After the long run, it's not just muscle soreness; my whole body feels heavy.
Quick intervals left me with a good burn in my quads, but I'm still moving well.
Resting heart rate's up, and so is the tightness in my hip flexors.
I went all out on the track; now my lower back and legs are paying for it.
Legs were heavy during the recovery run, especially in the Achilles area.
Worked my glutes, but they're holding up well.
Oddly sore considering the reduced mileage, might be nerves.
Feeling every mile today; my hip flexors and IT band are tight.
Based on the above observations and insights determined based on a trainee's demographics and training load (type of exercise, frequency and intensity).
Write a hypothetical muscle soreness responses by the trainee that is coherent with the above insights. If you refer to an exercise/workout, make sure it is discussed in the insights above.
The response should be a single sentence, as if the trainee is responding in an app on a mobile phone.

Supplementary Table 39 | Sleep disturbance survey. The survey is coded such that higher values correspond to greater sleep disturbance.

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
○	○	○	○	○

Supplementary Table 40 | Sleep impairment survey. The survey is coded such that higher values correspond to greater sleep impairment.

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
○	○	○	○	○

Supplementary Table 41 | 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

Supplementary Table 42 | 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 Supplementary Table 26 for an example)>

Sleep Summary:
<sleep summary (see Supplementary Table 27 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

Supplementary Table 43 | 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 Supplementary Table 27 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

Supplementary Table 44 | 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 Supplementary Table 27 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

Supplementary Table 45 | 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.

Supplementary Table 46 | 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 Supplementary Table 28 for an example)>
Today is <day of the week> the <Nth day>.

Here are some aggregate statistics for the last 30 days:
<aggregate statistics of daily activity metrics (see Supplementary Table 29 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

Supplementary Table 47 | 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 Supplementary Table 30 for an example)>
Today is <day of the week> the <Nth day>.

Here are some aggregate statistics for the last 30 days:
<aggregate statistics of sleep metrics (see Supplementary Table 31 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

Supplementary Table 48 | 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 Supplementary Table 32 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 Supplementary Table 33 for an example)>
Today is <day of the week> the <Nth day>.
Here are some aggregate statistics for the last 30 days:
<aggregate statistics of health metrics (see Supplementary Table 34 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

Supplementary Table 49 | 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

Supplementary Table 50 | 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: 52.0. heart rate variability root mean square of successive differences: 25.466. respiratory rate breaths per minute: 14.653. resting heart rate beats per minute: 68.8. awake minutes: 59.317. deep sleep minutes: 80.3. sleep duration minutes: 525516.688. efficiency: 0.886. overall sleep score: 85.018. percent of sleep in REM: 22.966. restlessness: 0.076. revitalization score: -1.0. sleep end time: -278.0. sleep start time: 546.467. sleep time minutes: 466.2. total number of minutes in awake state after sleep onset: 13.411. number of steps walked during the day: 6830.2. number of minutes spent in cardio zone during the day: 1.067. number of minutes spent in fat burn zone during the day: 20.867. number of minutes spent in peak zone during the day: 0.0. total multiplied minutes of exercise during the day: 23.0.
very restless: yes or no?

Supplementary Table 51 | 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 Methods and Supplementary Table 9. 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}

Supplementary Table 52 | AutoEval model training details by data mixture. For a given data mixture, "Training Samples" denotes the total number of training samples, "Max Steps" denotes the maximum number of training steps over the course of 20 epochs, and "Checkpoint Interval" denotes the number of steps between each checkpoint. The total number of steps varied per model since each training mixture comprised a different number of samples depending on the vertical and rater.

Mixture	Training Samples	Max Steps	Checkpoint Interval
Fitness			
All	7,138	4,462	56
Primary B (low variance)	2,141	1,339	16
Primary C (high variance)	1,836	1,148	14
Sleep			
All	4,850	3,032	38
Primary C (high variance)	1,004	628	8
Primary D (low variance)	1,256	785	10