nature portfolio

Corresponding author(s): Anastasiya Belyaeva, Xin Liu, Daniel McDuff, and Cory Y. McLean

Last updated by author(s): Jun 13, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All code provided with this manuscript is publicly available at github.com/google-health/consumer-health-research/tree/main/phllm. See the "Code Availability" section for details: "All code provided with this manuscript is available in the phllm folder of https://github.com/google-health/consumer-health-research. This repository includes code for interaction with the case study datasets, automatic evaluation of model responses to case study inputs, query and evaluation of multiple-choice question examinations, a reference implementation of the multimodal adapter for predicting patient-reported outcomes, and analysis and data visualization tools. The text LLM code underlying PH-LLM is Gemini Ultra 1.0."
Data analysis	All code provided with this manuscript is publicly available at github.com/google-health/consumer-health-research/tree/main/phllm. See the "Code Availability" section for details: "All code provided with this manuscript is available in the phllm folder of https://github.com/google-health/consumer-health-research. This repository includes code for interaction with the case study datasets, automatic evaluation of model responses to case study inputs, query and evaluation of multiple-choice question examinations, a reference implementation of the multimodal adapter for predicting patient-reported outcomes, and analysis and data visualization tools. The text LLM code underlying PH-LLM is Gemini Ultra 1.0."

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All input data and associated expert responses for all 857 case studies in sleep and fitness used for training and evaluation, along with all model responses to the 100 holdout case studies, are stored in JSON format and are publicly available at https://github.com/google-health/consumer-health-research/blob/main/phllm/data/sleep_case_studies.all.jsonl and https://github.com/google-health/consumer-health-research/blob/main/phllm/data/fitness_case_studies.all.jsonl, respectively. All input and target prediction data, which include bucketized age, daily sensor readings for 20 sensors over 15 days, and 16 binary survey responses, for all 4,163 samples in the PRO task are stored in JSON format. To ensure privacy of participants, data used for the PRO task are available to approved researchers for non-commercial research following registration and attestation of a data use agreement. Requests should be made to google-digital-wellbeing-dataset@google.com and will be evaluated on a weekly basis. Multiple choice questions are available at the listed URLs for sleep (https://www.boardvitals.com) and fitness (https://www.nsca.com/certification/cscs/certified-strength-and-conditioning-specialist-exam-description).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Self-reported gender was used in the sampling process for sleep case studies only to achieve a representative sample of age and gender as detailed in the Methods. Supplementary Table 25 shows the demographic statistics of the different datasets compared to 2021 US Census data and a large Fitbit user population. Our datasets were sampled from a broader Fitbit user population. While our sample is quite diverse, there are a few deviations when compared to the US Census data. Specifically, our sample skews with more women in the data used for evaluating the PH-LLM Patient-Reported Outcomes but with more men in the Fitness case studies. Our samples also skew older than the general population. These trends are not unexpected, for example see wearable age/gender trends here: https://connected-intelligence.com/about-us/press-releases/demographic-divide-fitness-trackers-and-smartwatches-attracting-very and the trends for physical activity here: https://www.cdc.gov/nchs/products/databriefs/db443.htm. Only self-reported gender is available in our datasets.
Reporting on race, ethnicity, or other socially relevant groupings	For our case study populations, because these were drawn from existing Fitbit/Pixel watch user populations, we only have age and gender information and not additional demographics.
Population characteristics	Supplementary Table 25 shows the demographic statistics of the different datasets compared to 2021 US Census data and a large Fitbit user population.
Recruitment	No recruitment was done for this retrospective study.
Ethics oversight	Ethics approval for the case studies was granted by Western Institutional Review Board-Copernicus Group IRB which classified the research as exempt under 45 CFR 46.104(d)(4). Ethics approval for the Digital Wellbeing Study was granted by the institutional review board of the University of Oregon.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We predetermined the target number of case studies given prior work on the number of samples required to successfully fine-tune a large language model (Zhou et al, LIMA: Less Is More for Alignment, arXiv:2305.11206, 2023). The number of multiple choice questions was dependent on the availability of examination questions available for our domains of interest. The number of PROs samples was dependent on the existing Digital Wellbeing Study (McDuff et al, The Google Health Digital Well-Being Study: Protocol for a Digital Device Use and Well-Being Study, JMIR Research Protocols, 13:e49189, 2024); we retained all samples that satisfied quality control and data missingness criteria outlined in the manuscript.
Data exclusions	Samples from the Digital Wellbeing Study were excluded using the quality control methods outlined in the manuscript to ensure data

Data exclusions	completeness. Exclusion criteria were not pre-established; these most stringent criteria were used once we observed that the dataset had many thousands of samples that satisfied these criteria.
Replication	Attempts at replicating model training and downstream analyses were successful (2 for AutoEval models, ~5-10 runs for MCQs, 2-3 for PROs).
Randomization	Samples were randomly split into train, test, validation, and holdout sets as described in detail in the Methods.
Blinding	Blinding is not relevant to data selection since this is a retrospective study without specific treatment arms.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*