nature portfolio

Corresponding author(s): Dylan Slack

Last updated by author(s): May 26, 2023

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	The TalkToModel system was used to collect data within the study. The TalkToModel system is provided in our open source package: https://github.com/dylan-slack/TalkToModel. TalkToModel relies on several software packages to produce results. These include dice-ml version 0.7.2, lime version 0.2.0.1, numpy version 1.21.6, pandas version 1.4.1, scikit-learn version 1.0.2, scipy version 1.8.0, shap version 0.40.0, and transformers version 4.17.0.
Data analysis	We used numpy version 1.21.6 and pandas version 1.4.1 and matplotlib version 3.5.1 to analyze the data in the study.

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

The German, Compas, and Diabetes datasets and models can be found at https://github.com/dylan-slack/TalkToModel/tree/main/data. The finetuned language models used for TalkToModel for each of these datasets can be found at https://huggingface.co/dslack/all-finetuned-ttm-models. The mturk generated dataset used to assess parsing accuracy and the accuracy results can be found at https://github.com/dylan-slack/TalkToModel/tree/main/experiments/parsing_accuracy. The user study response data is provided at https://github.com/dylan-slack/TalkToModel/blob/main/data/ttm-user-study-responses.csv.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	Sex and gender were not considered in the study design. The healthcare worker group sex information was split with 47% female 53% male participants (N=45). This information was self reported. We did not collect sex and gender information for the machine learning graduate student and professional group.
Population characteristics	See above.
Recruitment	We recruited the group of healthcare workers from the Prolific research service. Prolific provides an online interface to recruit high quality participants for user studies. Prolific enables us to filter participants by different characteristics, such as population characteristics and industry of occupation. Because we were interested in recruiting a group with healthcare experience, we included English speaking healthcare workers in our study. We informed the group of participants fitting our criteria provided by prolific the study would take around 30 minutes (we gauged this time based on pilot studies) and offered them $7.50 ($15/hour) to complete the study. We accepted the responses of the first 45 participants to complete the study. Note, participants were not allowed to begin the study if sufficient participants were already taking the study, to ensure work would not go uncompensated. We recruited the graduate and machine learning professional group from from machine learning Slack groups and email lists by advertising a paid study related to evaluating machine learning tools. Because individuals self report whether they are machine learning professionals, they may overestimate the extent of their expertise in machine learning, resulting in self-selection biases. For instance, there could be computer scientists with only a small amount of experience in machine learning that participate in the machine learning graduate student and professional group. This self-selection could impact the comparison between groups with differing levels of machine learning experience, because some individuals overestimate the scope of their expertise in machine learning. To try and ensure participants did not unduly self-select into the machine learning professional group, we specifically asked participants for the extent of their experience in machine learning by asking an equivalent in terms of educational attainment and planned to remove participants without graduate experience or higher (e.g., no experience/read articles online/undergraduate course/graduate course/published machine learning papers). We found our participants all stated they had graduate course experience or higher. Along the same lines, participants in the healthcare group may self-select into this group with occupations only tangentially related to healthcare, which could reduce the effects of the comparisons with this group. To help ensure that participants work in health care and control for self-selection biases in this regard, we asked them to state their occupation in the survey and found participants all had occupations directly related to healthcare (doctors, nurses, pharmacists, etc.). In general, we expect the measures we took in our study to help reduce adverse effects due to self-selection biases.
Ethics oversight	We received approval from the University of California, Irvine Institutional Review Board.

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

Behavioural & social sciences study design

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

Study description	The user study presented in the work includes mixed methods data, where we collected both quantitative metrics concerning accuracy and time taken solving a series of model understanding tasks with talktomodel compared to a dashboard style system. The specific metrics were the accuracy on the tasks (each user solved 10 tasks) and time taken per task (recorded as the time between question open and close). We additionally collected likert scores for several perceived quantitative metrics for each tasks, such as how confident they were in their answer, effort, usefulness of the interface, trust in the system, and likeliness to use again [1]. Each user rated these 5 metrics for each of the 10 questions. We additionally collected these same liker metrics at the end of the study, asking users to compare their preference between the two systems (e.g.,, whether they agree TalkToModel performed better than the dashboard, with respect to the likert metric). Finally, we collected qualitative feedback, where asked users to write a short description comparing their experience using both systems. [1] Quanze Chen, Tobias Schnabel, Besmira Nushi, and Saleema Amershi. Hint: Integration testing for ai-based features with humans in the loop. In International Conference on Intelligent User Interfaces. ACM, March 2022
Research sample	The research sample is a group of healthcare workers (N=45), consisting of 47% Female 53% Male, 20% Black, 29% White, 27% Mixed 14% Other, average 26.5 years old (min 20, max 67). This sample is not guaranteed to be representative. The other group recruited in the study is machine learning graduate students and professionals. We did not collect covariate information for this group. This group is not guaranteed to be representative. These two groups were appropriate for our study because we wished to analyze the effectiveness of TalkToModel for groups with varying skills and expertise in machine learning.
Sampling strategy	The populations were sampled with convenience sampling. We determined the sample sizes need to achieve statistical significance by following prior related studies within machine learning and natural language processing communities that evaluate the performance of machine learning systems with end users and that achieved statistical significance. We used similar sample sizes with our human evaluation studies to achieve statistical significance [1, 2, 3]. For instance, our study with graduate students had 13 participants, where the related works [1] had 18 participants and [2] had 13 participants. Similarly, we used 45 participants within the health care worker group, which is again comparable to prior works on studying the machine learning systems with end users, e.g., [3] which uses 31 participants. Because TalkToModel shares many similar characteristics with these prior works in machine learning and natural language processing that achieved statistical significance (e.g., evaluating performance of a machine learning system with end users such as machine learning graduate students), using comparable sample sizes to those presented in the related literature is an appropriate choice for our work. [1] Beyond Accuracy: Behavioral Testing of NLP models with CheckList. Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, Sameer Singh Association for Computational Linguistics (ACL), 2020 [2] Quanze Chen, Tobias Schnabel, Besmira Nushi, and Saleema Amershi. Hint: Integration testing for ai-based features with humans in the loop. In International Conference on Intelligent User Interfaces. ACM, March 2022 [3] Dylan Slack, Sophie Hilgard, Sameer Singh, Himabindu Lakkaraju. Reliable Post hoc Explanations: Modeling Uncertainty in Explainability. NeurIPS, 2021.
Data collection	We sent out an online survey using Qualtrics. Results were stored in a qualtrics database. We did not observe the participants taking the study. The researchers were blinded to the experimental conditions (the tasks and order of the machine learning systems given to the study participants were randomized). The study hypothesis was not blind to the researchers.
Timing	We started collection on June 12th, 2022 and closed collection on June 23rd, 2022.
Data exclusions	We excluded 1 data point for a healthcare worker who failed a sanity check question in our survey. We included sanity check questions as a pre-established exclusion criteria, to ensure we received data from participants who were actively paying attention to the survey.
Non-participation	There were 10 participants who dropped out, either by opening the survey and timing out, or closing the survey tab, having not returned responses.
Randomization	The ordering of the tasks and sections of the survey were randomized.

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

Methods

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