Peer Community In Registered Reports

Using large language models to predict relationships among survey scales and items

*A recommendation by **Matti Vuorre** based on peer reviews by **Johannes Breuer**, **Hu Chuan-Peng** and **Zak Hussain** of the STAGE 2 REPORT:*

Björn E. Hommel, Ruben C. Arslan (2025) Language models accurately infer correlations between psychological items and scales from text alone. OSF, ver. 4, peer-reviewed and recommended by Peer Community in Registered Reports.
https://osf.io/preprints/psyarxiv/kjuce_v4

Submitted: 03 February 2025, Recommended: 28 April 2025

Cite this recommendation as:
Vuorre, M. (2025) Using large language models to predict relationships among survey scales and items. *Peer Community in Registered Reports*, 100990. 10.24072/pci.rr.100990

Published: 28 April 2025

How are the thousands of existing, and yet to be created, psychological measurement instruments related, and how reliable are they? Here, Hommel and Arslan (2025) trained a language model–SurveyBot3000–to provide answers to these questions efficiently and without human intervention. In their Stage 1 submission, the authors described the training and pilot validation of a statistical model whose inputs are psychological measurement items or scales, and outputs are the interrelationships between the items and scales, and their reliabilities. The pilot results were promising: SurveyBot3000's predicted inter-scale correlations were strongly associated with empirical correlations from existing human data. In this Stage 2 report, the authors further examined the model's performance and validity. In accordance with their Stage 1 plans, they collected new data from 450 demographically diverse participants, and tested the model's performance fully out of sample. The model's item-to-item correlations correlated at r=.59 with corresponding item-to-item correlations from human participants. The scale-to-scale correlations were even more accurate at r=.83, indicating reasonable performance. Nevertheless, the authors remain justifiably cautious in their recommendation that the "synthetic estimates can guide an investigation but need to be followed up by human researchers with human data." The authors documented all deviations between Stage 2 execution and Stage 1 plans in extensive online supplements. These supplements also addressed other potential issues, such as the potential for data leakage (finding the results in the training data) and robustness of results across different exclusion criteria. The authors' proposed psychometric approach and tool, which is freely available as an online app, could prove valuable for researchers either looking to use or adapt existing scales or items, or when developing new scales

or items. More generally, these results add to the growing literature on human-AI research collaboration and highlight a practical application of these tools that remain novel to many researchers in the field. As such, this Stage 2 report and SurveyBot3000 promise to contribute positively to the field. The Stage 2 report was evaluated by two reviewers who also reviewed the Stage 1 report, and a new expert in the field. On aggregate, the reviewers' comments were helpful but relatively minor; the authors improved their work in a resubmission, and the recommender judged accordingly that the manuscript met the Stage 2 criteria for recommendation. **URL to the preregistered Stage 1 protocol:** https://osf.io/2c8hf **Level of bias control achieved:** Level 6. *No part of the data or evidence that was used to answer the research question was generated until after IPA.* **List of eligible PCI RR-friendly journals:**

- Advances in Methods and Practices in Psychological Science
- Collabra: Psychology
- International Review of Social Psychology
- Peer Community Journal
- PeerJ
- Personality Science
- Royal Society Open Science
- Social Psychological Bulletin
- Studia Psychologica
- Swiss Psychology Open

References:

Hommel, B. E. & Arslan, R. C. (2025). Language models accurately infer correlations between psychological items and scales from text alone [Stage 2]. Acceptance of Version 4 by Peer Community in Registered Reports. https://doi.org/10.31234/osf.io/kjuce_v4

Reviews

Evaluation round #1

DOI or URL of the preprint: **https://osf.io/preprints/psyarxiv/kjuce_v3**
Version of the preprint: 3

Authors' reply, 04 April 2025

Dear Dr. Vuorre,

We are grateful for the opportunity to submit our revision for our Stage 2 Registered Report titled "Language models accurately infer correlations between psychological items and scales from text alone" for review at PCI: RR. We again extend our sincere appreciation to the three reviewers for their thoughtful feedback.

In this revision, we have addressed all the points raised by you and the reviewers. We have corrected the language to reflect past tense throughout the manuscript and drafted a Competing Interests Statement that hopefully clarifies that no conflicting interests exist that impact our work on the manuscript. The online repository now contains an updated readme and all the necessary files to reproduce our findings. We have also refined our discussion to accommodate recent work by Wulff & Mata (2025) as well as Schoenegger et al. (2025). In addition to these points, we have conducted an extensive search of the data used in model pre-training of the SBERT (i.e., the base) model to verify that our high accuracy is not confounded by data leakage. Details on this investigation can be found in the updated supplemental section.

We confirm that no data generated as part of this manuscript are currently under review or published elsewhere. Additionally, this manuscript is original, not previously published, and not under concurrent consideration elsewhere.

Sincerely,

Björn E. Hommel and Ruben C. Arslan
Department of Personality Psychology and Psychological Assessment, University of Leipzig

Download author's reply

Download tracked changes file

Decision by Matti Vuorre, posted 12 March 2025, validated 12 March 2025

Revision invited

Dear Authors,

Thank you for your Stage 2 submission. Two reviewers from Stage 1 returned, and I invited an additional reviewer to replace a third reviewer from Stage 1 who wasn't able to evaluate the Stage 2 report. Overall the three reviewers commented positively on the submission, but pointed out somewhat minor but still important areas of improvement in your write-up.

I am happy to invite you to submit a revision & response letter, and anticipate a quick turnaround afterwards. In your revision/response, please address the reviewers' comments while paying attention especially to the following points:

1. Ensure the tense is consistent & appropriate throughout (e.g. in Intro, the validation study is now done and so you no longer "plan to validate"; "All necessary support is in place for the proposed research." is now redundant, etc.)
2. Include an appropriate related interests statement or similar
3. Overall the transparency of the document and associated code/programs deserves praise, but as noted there are some areas where you should update your code/documentation as noted by HC-P
4. Update discussions of Wulff & Mata and pre-trained model concerns as noted by ZH
5. Are you able to better link to the supplementary online materials? e.g. there are references to "Supplementary Notes" but I had to dig around in the OSF repo to find the correct file(s). I understand this may be a limitation of the OSF but perhaps you can see if this could be improved (with e.g. directly linking to files' persistent DOIs.)

Kind regards,
Matti Vuorre

Reviewed by Johannes Breuer, 07 March 2025

As for the stage 1 version of this paper, I have enjoyed reading the stage 2 report and believe that this can make a valuable and impacful contribution to different fields (esp. psychometrics/scale development and

research on the use of AI/LLMs for science).

The authors have followed the preregistered procedures laid out in the stage 1 report. The few minor deviations (e.g., in the sampling/inclusion criteria for respondents or the sample size) are transparently reported.

I only have a couple of minor remarks that should be relatively easy to address for the authors. I will list those in chronological order in the following.

I would suggest changing the language in the Introduction (and elsewhere) describing the validation study from future tense to past tense (or present tense) as it has now been conducted.

On p. 18, 2nd para, the upper bound of the CI for the manifest correlation for the accuracy of synthetic scale correlations is missing. It appears that the formatting of the whole parentheses in which this is contained is not correct/broken.

A minor note related to Figure 5: From a data viz perspective, I prefer it if the y-axis and the bars start at 0 (i.e., without any whitespace between the bars and the x-axis).

Regarding Figure 6: The Figure distinguishes between "latent outcomes (SEM)" and "manifest". This distinction appears for the first time in this form/wording here. Hence, I think that this should be (further) explained in the paper before the figure is presented.

In the Discussion, maybe the parts addressing the general/broader context (and the implications of the presented findings for this) could be extended a bit to elaborate further on what the results (can) mean for research practices in psychology and other social and behavioral sciences (see, e.g., the recent paper by Binz et al., 2025).

Literature cited in this review

Binz, M., Alaniz, S., Roskies, A., Aczel, B., Bergstrom, C. T., Allen, C., Schad, D., Wulff, D., West, J. D., Zhang, Q., Shiffrin, R. M., Gershman, S. J., Popov, V., Bender, E. M., Marelli, M., Botvinick, M. M., Akata, Z., & Schulz, E. (2025). How should the advancement of large language models affect the practice of science? Proceedings of the National Academy of Sciences, 122(5), e2401227121. https://doi.org/10.1073/pnas.2401227121

Reviewed by Hu Chuan-Peng, 10 March 2025

I am pleased to review this Stage 2 Registered Report. Overall, the work is rigorous, transparently documented, and addresses its objectives effectively. The results demonstrate the model's ability to mirror real-world reliability coefficients, scale correlations, and covariance patterns. Please see below for my revision suggestions to strengthen the manuscript further:

(1) Analytical reproducibility concerns

The GitHub repository (https://github.com/synth-science/surveybot3000), which was linked to OSF, requires updates to align with the Stage 2 submission. For example, the repository's README files appear outdated and do not reflect the current workflow or documentation; Critical files/folders necessary for full reproducibility are missing (e.g., ../synth-rep-dataset/, ignore/, docs/, ignore.random_scales_rr.rds). If these omissions are intentional (e.g., due to privacy constraints), this should be explicitly justified in the repository and manuscript.

(2) Clarify how accuracy was estimated in methods

The manuscript should briefly summarize how prediction accuracy was quantified (e.g., error metrics, validation procedures). While Supplementary Note 1 mentions two approaches, the main text lacks sufficient detail for readers to evaluate methodological rigor. Please clarify this in the Methods section.

(3) Discussion of figure 4

The variation in prediction error across prediction types (Figure 4) is intriguing but not mentioned in the discussion. Could these patterns reflect meaningful differences in model performance? Consider addressing this briefly in Discussion.

(4) Conflict of interest statement

A conflict of interest statement is absent. Given the first author's affiliation with magnolia psychometrics GmbH (a commercial entity), please confirm whether this paper's recommendations or outcomes could be perceived as benefiting their commercial interests.

Signed
Hu Chuan-Peng
(I have signed the Peer Reviewers' Openness Initiative, https://www.opennessinitiative.org)

Reviewed by Zak Hussain, 12 March 2025

An important and timely contribution - I look forward to seeing this work published! I would, however, suggest the following (minor) revisions:

- I would clarify in the abstract which 'out-of-sample accuracy' metric is being used (Pearson correlation, coefficient of determination etc.).
- Perhaps consider re-wording 'out-of-sample accuracy' to 'out-of-sample performance'. The term 'accuracy' often suggests that the task being performed is a classification task, which is not the case here.

"In a process called fine-tuning, the model then retains its originally learned weights but learns to carry out a specific task, such as text classification. Essentially, the model builds on the fundamental knowledge acquired during pre-training to adapt to specialised tasks, even with limited training data. This concept is known as few-shot learning."

- This paragraph appears to blur the distinction between related but distinct concepts. In the context of language modeling, fine-tuning is typically understood as updating at least some model weights for a specific task. The phrase "the model then retains its originally learned weights" could therefore be misleading unless referring to methods where weights remain frozen (e.g., in-context learning, of which few-shot learning is a special case). If the authors intended to describe an approach where the weights remain frozen, I would suggest using "in-context learning" rather than "fine-tuning". However, given that the research later performs full-model fine-tuning (i.e., with weight updates), I assume this is what the authors mean. In that case, it would be helpful to specify the distinction more explicitly to avoid conflating fine-tuning with few-shot learning, which often relies on frozen weights and contextual adaptation rather than explicit weight modification (see, e.g., Brown et al., 2020). In this case, I would recommend sticking to the term "transfer learning" instead of "few-shot learning".

"However, their [Wulff & Mata, 2023] approach relied on pre-trained models that were not adapted to the domain of survey items and do not appreciate that empirical item correlations are often negative because of negation."

- Although earlier version of Wulff & Mata's work do not adapt their model to the domain of the survey items, the most recent versions does (Wulff & Mata, 2025).
- As a non-expert in psychometrics, it is not immediately clear to me what the value is of being able to predict polarity/negative correlations. Perhaps the value of doing this could be made clearer somewhere early in the paper, especially since it appears to be a key contribution of the work. In particular, it would be useful to have an explanation of how predicting polarity might help researchers 'evaluate new measures against existing scales, reduce redundancy in measurement, and work towards a more unified behavioural science taxonomy', as expressed in the abstract as a main contribution of the work.
- In general, I would find it helpful if the authors expanded on the main differences/contributions of their work relative to Wulff & Mata (2025). I appreciate that these two pieces of work were carried out in parallel, and thus some overlap is inevitable. However, I get the impression that differences/contributions of the present work could be emphasized more clearly.

"Because OpenAI's large language models obtain knowledge from scraping large quantities of internet text, they presumably have seen items from existing measures co-occur in online studies and public item repositories."

- This is an important consideration. However, it is not obvious to me that the pre-trained model used (all-mpnet-base-v2) is free from such concerns. Perhaps the authors could expand on the extent to which they believe information on their survey items might have leaked into the datasets (Wikipedia, BooksCorpus, OpenWebText, CC-News, and Stories) used to pre-train all-mpnet-base-v2.

"In other words, our fine-tuned LLM explained 80% of the latent variance in scale intercorrelations, based on nothing but semantic information contained in the items."

- It may not be entirely accurate to say that the model's predictions are based **only** on the **semantic information** contained in the items. The authors already noted that pre-trained models could be leveraging other sorts of information to make their predictions, such as co-occurence patterns of the items found "in online studies and public item repositories". Even smaller language models like all-mpnet-base-v2 likely encode more than just word meanings. For instance, such models have been shown to predict response times and reading difficulty, which can be impacted by other factors such as item complexity, social desirability, or phonetics.
- In general, more thorough discussion of what the model could be leveraging to predict item similarity would benefit this work, especially since it has implications for a possible upper bound on model performance when predicting item correlations. This could also be informed by speculation/research on the extent to which observed item correlations reflect semantic similarity versus other factors.

References:

- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.
- Wulff, D.U., Mata, R. Semantic embeddings reveal and address taxonomic incommensurability in psychological measurement. Nat Hum Behav (2025). https://doi.org/10.1038/s41562-024-02089-y