


Language models accurately infer correlations between psychological items and scales from text alone

Björn E. Hommel^{1, 2*} and Ruben C. Arslan¹

¹ Wilhelm Wundt Institute of Psychology, Leipzig University, Germany

² magnolia psychometrics GmbH, Germany

Björn E. Hommel  <https://orcid.org/0000-0002-7375-006X>

Ruben C. Arslan  <https://orcid.org/0000-0002-6670-5658>

* Corresponding author: Björn E. Hommel (bjoern.hommel@uni-leipzig.de)

Björn E. Hommel and Ruben C. Arslan contributed equally to this paper.

Online supplement:

- Stage 1 preprint: https://osf.io/preprints/psyarxiv/kjuce_v1
- Statistical reports and interactive plots: <https://synth-science.github.io/surveybot3000/>
- OSF (Code & Data): <https://osf.io/z47qs/>,
- App: <https://huggingface.co/spaces/magnolia-psychometrics/synthetic-correlations>

Abstract:

Many behavioural scientists do not agree on core constructs and how they should be measured. Different literatures measure related constructs, but the connections are not always obvious to readers and meta-analysts. Many measures in behavioural science are based on agreement with survey items. Because these items are sentences, computerised language models can make connections between disparate measures and constructs and help researchers regain an overview over the rapidly growing, fragmented literature. Our fine-tuned language model, the SurveyBot3000, accurately predicts the correlations between survey items, the reliability of aggregated measurement scales, and intercorrelations between scales from item positions in semantic vector space. **We measured the model's performance as the convergence between its synthetic model estimates and empirical coefficients observed in human data.** In our pilot study, **this** out-of-sample accuracy for item correlations was .71, .89 for reliabilities, and .89 for scale correlations. In our preregistered validation study using novel items, the out-of-sample accuracy was slightly reduced to .59 for

item correlations, .84 for reliabilities, and .84 for scale correlations. The synthetic item correlations showed an average prediction error of .17, with larger errors for middling correlations. Predictions exhibited generalizability beyond the training data and across various domains, with some variability in accuracy. Our work shows language models can reliably predict psychometric relationships between survey items, enabling researchers to evaluate new measures against existing scales, reduce redundancy in measurement, and work towards a more unified behavioural science taxonomy.

Introduction

Behavioural science struggles to be cumulative in part because scientists in many fields fail to agree on core constructs (Bainbridge et al., 2022; Sharp et al., 2023). The literature silos, which consequently develop, can appear unconnected but pursue the same phenomena under different labels (see e.g., grit and conscientiousness; Credé et al., 2017).

One reason why connections are lacking is the asymmetry inherent in measure and construct validation: adding novel constructs to the pile is easier than sorting through it. Investigators can easily invent a new ad-hoc measure and benefit reputationally if a new construct becomes associated with their name (Elson et al., 2023; Flake & Fried, 2020). By contrast, finding out whether a purported new construct or measure is redundant with the thousands of existing ones is cumbersome and can cause conflict with other researchers (Bainbridge et al., 2022; Elson et al., 2023). The same holds for replicating construct validation studies and reporting evidence of overfitting or other problems (Hussey et al., 2024; Kopalle & Lehmann, 1997).

Untangling the "nomological net"—a term coined by Cronbach and Meehl (1955) to describe the relationships between measures and constructs—has become increasingly difficult given the growing number of published measures (Anvari et al., 2024; Elson et al., 2023).

Conventional construct validation methods, though effective in mapping these relationships, do not scale to, for instance, the thousands of measures that might be related to neuroticism. To tackle this problem, Condon and Revelle (2015; see also Condon, 2017; Condon et al., 2017) have championed the Synthetic Aperture Personality Assessment in which survey participants respond to a small random selection of a large set of items from the personality literature. Over time, as the sample size grows, this procedure allows estimating pairwise correlations between all items. Although the approach is efficient, each new item requires thousands of participants to answer the survey before it can be correlated with all existing items. Hence, the approach cannot be used to quickly evaluate new proposed scales. What is missing is an efficient way to prioritise, to prune the growth in constructs and measures and to sort through the disorganised pile of existing measures.

Natural language processing could provide this efficiency. In the social and behavioural sciences, subjective self-reports are one of the predominant forms of measurement. The textual nature of survey items lends itself to natural language processing. Recently, transformer models have become the state-of-the-art in language models (Vaswani et al., 2017), displaying proficiency in understanding and generating text. They have dramatically reduced the costs of many tasks and chores, notably in programming and generating images from verbal prompts. Although capabilities for natural language generation are currently more visible in the public eye through the use of chat-like interfaces, they are backed by capabilities in natural language understanding (e.g., classifying or extracting features from text).

On a technical level, this understanding is implemented by the so-called encoder block, which processes input text and encodes it as a high-dimensional numeric vector. The vector representation of a word like "party" in the resulting semantic vector space is context-

dependent. The same word will yield a different vector representation if it occurs in the statement “I am the life of the party” compared to “I always vote for the same party”. The encoder block's ability to contextualise words is crucial for recognizing the nuances of language. At heart, the efficiency of the transformer model can largely be attributed to its self-attention mechanism (Vaswani et al., 2017). As the name suggests, it is loosely analogous to the executive function in human cognition. Instead of “memorising” an entire corpus of text, word by word, the attention mechanism weights the relevance of words in a context window for a given target word.

Transformer models excel in transfer learning, that is, they adapt to new tasks easily (Tunstall et al., 2022). Following the pre-training stage, which establishes a base level of linguistic expertise, the models can undergo domain adaptation, which involves training the model on a corpus of text specifically curated for the task at hand. In a process called fine-tuning, the model then **learns to carry out a specific task (e.g., text classification). Fine-tuning often involves slight architectural adjustments to the model's output layer, although the term is used somewhat inconsistently in the literature to describe various adaptation approaches.** Essentially, the model builds on the fundamental knowledge acquired during pre-training to adapt to specialised tasks, even with limited training data. High-quality annotated training data is key for the domain adaptation that turns generalists into specialists.

Using linguistic information to scaffold scientific models has a long history in personality psychology, where the lexical hypothesis states that more important personality characteristics are more likely to be encoded as words. To find important personality dimensions, researchers had human subjects rate themselves on prominent adjectives, or *items*, identified systematic correlations between items, and applied factor analytic techniques to reduce the number of dimensions. The most popular organising framework, the Big Five, was distilled from personality-descriptive items in this manner (Digman, 1990).

Pre-transformer era attempts to use semantic features of items to predict associations between measurement scales using latent semantic analysis have demonstrated moderate utility (Arnulf et al., 2014; Larsen & Bong, 2016; Rosenbusch et al., 2020). As the ability of computerised language models to capture meaning has grown, researchers have sought to directly quantify relationships between adjectives from textual data (Cutler & Condon, 2022), to assign items to constructs (Fyffe et al., 2024; Guenole et al., 2024), to directly predict item responses (Abdurahman et al., 2024; Argyle et al., 2023) and quantify answers to open-ended questions (Kjell et al., 2019, 2024).

Wulff & Mata (2023) used large language models (LLMs) to map survey items to vector space and predict empirical item correlations. They tested various transformer models for their ability to predict properties of psychological inventories. They observed a correlation of $r = .22$ between the semantic similarities of items as judged by OpenAI's ada-002 model (Greene et al., 2022) and the item correlations estimated in empirical data, with accuracy improving when aggregating vectors to the scale level. Their work shows large language models can approximately infer item correlations and outperform latent semantic analysis. However, their 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. This approach cannot be expected to unlock the latent ability of the models, but rather to give a lower bound of their usefulness. At the same time, pre-trained models can overfit to their training data. 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 (see [Supplementary Note 11 for details on training data leakage](#)). The results for survey items that inadvertently were part of the training data can lead to more optimistic results than could be expected for novel items.

We have adapted a sentence transformer model to the domain of survey response patterns and trained our model, the SurveyBot3000, to place items in vector space. The distances between item pairs in vector space produce what we will call *synthetic* item correlations, scale correlations, and reliabilities. These synthetic estimates can potentially help to cheaply evaluate measures and constructs. [We have validated that the SurveyBot3000 can approximately infer empirical item correlations beyond its training data, by preregistering the model's synthetic estimates before collecting empirical data.](#) Based on our pilot study, we predicted that the model will exhibit substantial accuracy in inferring empirical item correlations ($r = .71$, 95% CI [.70;.72]), and even higher accuracy in inferring latent correlations between scales ($r = .89$ [.88;.90]) and in inferring reliability coefficients ($r = .89$ [.84;.94]). We detail our predictions in our Design Table.

Our model can be put to work in multiple areas. Synthetic correlations will always require careful follow-up with empirical data, but they can be used to search and prioritise. Authors can use our model as a semantic search engine to find existing constructs and measures and avoid reinventions. Synthetic correlations could be used as inputs for more realistic *a priori* power analyses. Scientific reviewers can use it to flag optimistic reliability coefficients and unstable factor structures, especially when researchers have not validated an ad-hoc measure out-of-sample yet. Generally, discrepancies between reported estimates and LLM-based synthetic estimates can motivate greater attention to replication and construct validation. Finally, meta-scientists and measurement researchers can use the model to start sorting through the pile of tens of thousands existing constructs and measures (Anvari et al., 2024; Elson et al., 2023).

As a showcase, we have made the model available as an app on Huggingface. Researchers can enter item texts and the app will generate synthetic item correlations, scale correlations and reliability coefficients. The app contains a prominent cautionary note to discourage researchers from taking the synthetic estimates at face value before further validation has occurred.

Methods

Materials, data, and code for the present study are available through the Open Science Framework: <https://osf.io/z47qs/>. Data pre-processing, model training, and statistical analyses were conducted using Python (version 3.10.12; Van Rossum & Drake, 2009), R (version

4.3.1; R Core Team, 2023), with an Nvidia GeForce RTX 2070 Super GPU, using the CUDA 11.7.1 toolkit (NVIDIA et al., 2022).

Ethics information

The planned research complies with the ethics guidelines by the German Society for Psychology (Berufsverband Deutscher Psychologinnen und Psychologen, 2022). Data used in model training were collected by third parties, as shown in the online supplemental section (<https://osf.io/z47qs/>). Participants in the validation study **were** recruited from the crowdsourcing platform *Prolific*, and compensated at a median wage of \$12 per hour. Informed consent **has been** obtained from all human respondents. Ethics approval for the validation study has been granted from the Institutional Review Board (IRB) at Leipzig University. All necessary support is in place for the proposed research.

Pre-trained language model

Our preliminary work has focused on improving the predictions of item correlations with sentence transformer models using high-quality training corpora for domain adaptation. We modified a LLM to generate synthetic item correlations by fine-tuning a pre-trained sentence transformer model (Reimers & Gurevych, 2019). Unlike conventional transformer models used in natural language understanding tasks which produce vector representations of individual tokens (i.e., basic linguistic units, such as words or syllables), sentence transformers produce vector representations for longer sequences of text (e.g., sentences).

Sentence transformers—specifically the bi-encoder architecture used throughout this research—work by using two parallel LLMs that process text inputs independently but share the same structure and parameters. The central idea behind these models is to capture the semantic essence of a sentence. One method to accomplish this is by pooling (e.g., averaging) the contextualised token vectors for each of the two models and then combining them. The underlying neural network then learns these combined representations by predicting sentence similarities, for instance using natural language inference data. In natural language inference, a given text (i.e., the premise) is evaluated based on its relation to another text (i.e., the hypothesis), classified as either contradicting, entailing, or being neutral to it. The network's output layer consists of three neurons, each representing one of these classes. The model's learning effectiveness is assessed using cross-entropy loss, with improvements in sentence vector representation achieved through backpropagation. Interested readers are referred to Reimers & Gurevych (2019), as well as Schroff et al. (2015) for further details on bi-encoders. Accessible in-depth introductions to transformer models and deep neural networks can be found in Hussain et al. (2023) and Hommel et al. (2022).

We chose the *all-mpnet-base-v2* model (hereafter referred to as the “SBERT model” for further fine-tuning from the Hugging Face model hub (*Hugging Face Model Hub*, n.d.), based on its commendable performance across 14 benchmark datasets (*Pretrained Models — Sentence-Transformers Documentation*, n.d.). This pre-trained model is a sentence-

transformer adaptation of the *mpnet-base* model (Song et al., 2020), initially trained on 160 gigabytes of English language text, including Wikipedia, BooksCorpus, OpenWebText, CC-News, and Stories. The SBERT model places sentences in a 768-dimensional semantic vector space. Distances in this Euclidean space can be computed using, for instance, cosine similarity. In our case, we hypothesised that the cosine similarity between the vector representations of any two survey items (e.g., personality statements) should correspond to the correlation coefficients obtained from survey data.

Domain adaptation and fine-tuning

We fine-tuned the pre-trained model in two steps. In the first step, we trained the model to distinguish between semantically opposing concepts. In the second step, we trained the model to predict pairwise item correlations, using survey data. Figure 1 depicts the multi-step training procedure.

Step 1: Polarity calibration Although cosine similarity spans from -1 to 1, negative coefficients are rarely produced when comparing vector representations of sentences (cf. the croissant shape of the top left plot in Figure 2). This limitation primarily arises because the high-dimensional vector representation of sentences encodes a range of abstract linguistic features, many of which tend to be positively correlated across text sequences. This poses a challenge in accurately predicting correlations for items of opposing scale polarities, such as those on the introversion-extraversion continuum. To illustrate, when assessing cosine similarity between items from the pre-trained model, the item “I am the life of the party” produces comparable coefficients with “I make friends easily” ($\theta = .32$) and “I keep in the background” ($\theta = .35$). This occurs even though the last item represents the polar opposite of the first item.

We fine-tuned the pre-trained model with the goal of maximising the cosine distance between vector representations of opposing concepts. We achieved this by augmenting the Stanford Natural Language Inference corpus (SNLI version 1.0, see also Supplementary Note 3; Williams et al., 2018) for our purposes. SNLI comprises around 570,000 sentence pairs, each labelled for textual entailment as either contradiction, neutral, or entailment. We re-labelled each sentence pair by additionally assigning a magnitude to the semantic relationship. We let the pre-trained SBERT model generate the cosine similarity of the sentence pair (e.g., “the moon is shining” and “it is a sunny day”, $\theta = .46$), but assigned a negative direction if the pair was labelled as contradictory (e.g., $\theta = -.46$). Hence, our new criterion combined the magnitude and direction of the similarity, capturing various forms of negation in the process. The fine-tuned model was then trained to predict this new criterion, so that it would learn that similar sentences have negative cosine similarities when one sentence negates or contradicts the other (see Supplementary Note 6 for more detailed evaluation metrics).

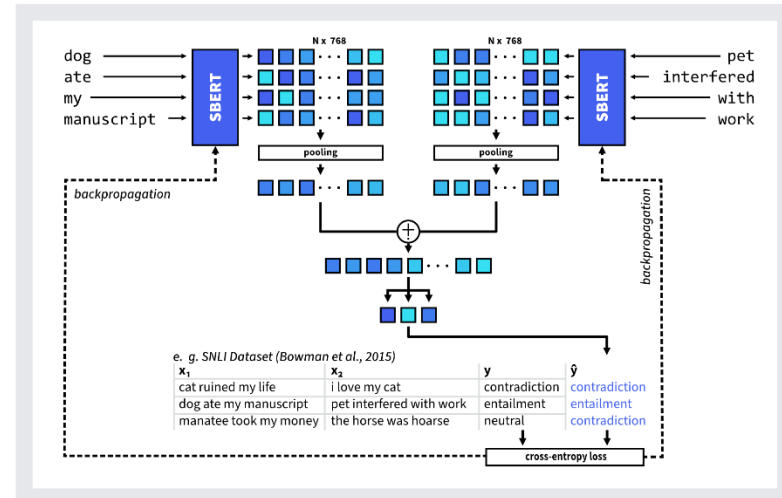
Step 2: Domain adaptation We found that the SBERT model's predictions of item correlations were skewed by the presence of non-trait-related text in the item stems. Specifically, we identified a tendency for item correlations to be overestimated in statements

containing the same adverbs of frequency. For example, the phrase “I *often* feel blue” from the depression facet of the NEO-PI-R in the IPIP exhibits similar cosine similarity to the two items “I feel that my life lacks direction” ($\Theta = .28$) and “I *often* forget to put things back in their proper place” ($\Theta = .26$), even though the first item is also from the depression facet while the second is from the orderliness facet.

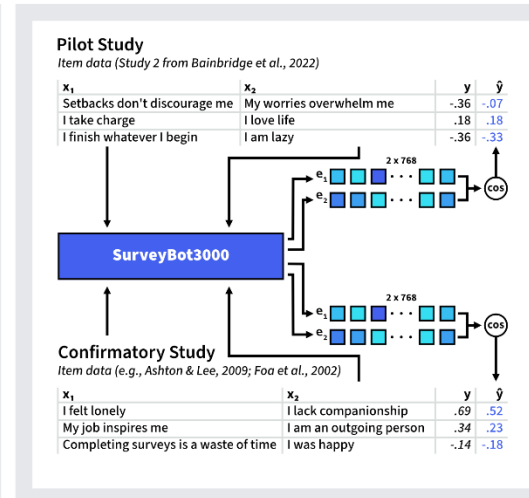
To address this, we aimed to fine-tune the model to focus on text segments that convey information relevant to psychological traits and their similarity. This adjustment aimed to enhance the model's accuracy in identifying and processing trait-relevant language and to teach it about personality structure, thus improving the validity of its synthetic correlations. We compiled training data from 29 publicly available online repositories (see Supplementary Note 4). Our inclusion criteria for the corpus mandated that raw item-level data be available, a minimum sample size of $N \geq 300$, the use of a rating scale as response format, and clear mapping of item stems to variable names in the datasets. In pre-processing, we retained pairwise Pearson coefficients from the lower triangular matrix across all datasets and cleaned and standardised item stems. Further details on the preprocessing of data can be found on the OSF (<https://osf.io/bfhzy>). For cross-validation purposes, we distributed each item pair among training, validation, and test partitions, adhering to an 80-10-10 split. To avoid overfitting, we ensured that all items were unique to their partition. This led to the exclusion of a substantial portion of our training data. Specifically, from the initial pool of 204,424 item pairs, we retained 90,424 pairs. Of these, we randomly allocated 74,339 pairs (82%) to the training partition, 6,832 pairs (8%) to the validation partition, and 9,253 pairs (10%) to the test partition. To mitigate the risk of the model learning idiosyncratic characteristics inherent to the dataset—item stems within a dataset are more likely to exhibit resemblance than between datasets—we used an additional holdout dataset. This dataset comprised 87,153 item pairs obtained from Bainbridge et al. (2022) thereby providing a robust measure for evaluating the model's generalizability to novel English language items about personality and related individual differences. To ensure the integrity of the holdout dataset, any items not exclusive to it were eliminated from the training, validation, and test partitions.

We optimised the hyperparameters for fine-tuning the model using the Optuna library in Python (version 3.1.1; Akiba et al., 2019), with a focus on enhancing the model's ability in predicting item correlations within the test partition. Details of the final hyperparameter selection are available in the online supplemental material (<https://osf.io/b5ua7>).

a) Pretraining — Base Model (SBERT)



c) Validation



b) Finetuning — SurveyBot3000

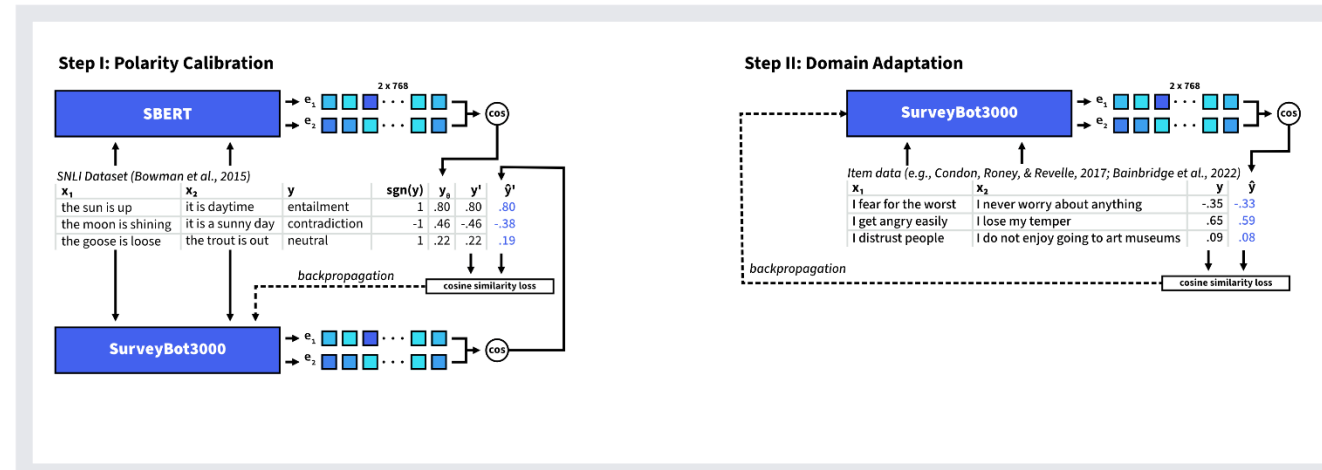


Figure 1. Multi-step training procedure for the SurveyBot3000, which produces synthetic estimates of inter-item correlations.

Pilot study

We found that the SurveyBot3000 model was highly accurate for all partitions of the *curated* corpus. Empirical inter-item correlations and synthetic correlations were accurately predicted in the test set $r = .69$ ($df = 9,251$; 95% CI [.67, .70]) and in the validation set $r = .71$ ($df = 6,830$; 95% CI [.70, .72]). That accuracy was high in both test and validation set shows the model's strong generalizability within the corpus.

The SurveyBot3000 model was then tested using 87,153 item pairs obtained from Bainbridge et al. (2022), the holdout dataset we withheld from the training process to avoid over-fitting. Adjusted for sampling error in the empirical data (see Supplementary Note 1), the model's synthetic correlations predicted the empirical inter-item correlations with an accuracy of $r = .71$ (95% CI [.70;.72], manifest correlation $r = .67$ [.67; .68], Figure 2). This consistency with the test-set performance shows the model's ability to generalise beyond the idiosyncratic properties of the data seen in training. Figure 2 shows the prediction of item correlations through semantic similarity, as estimated by the SBERT and SurveyBot3000 models. The SBERT model had substantially lower accuracy in predicting inter-item correlations in our holdout (manifest $r = .19$ [.18;.19]).

We further investigated the model's ability to predict scale reliabilities, which can be calculated from inter-item correlation matrices. Given that scales are typically designed to exhibit high internal consistency, we observed limited variability in the internal consistency measures across the 107 scales and subscales in the holdout dataset. Empirical Cronbach's alpha values had a mean of .75 ($SD = .10$) and ranged from .35 to .93. When new scales are designed, reliability varies more widely. We therefore circumvented the problem of restricted variance by randomly sampling items to create 200 additional, varied scales. We found that synthetic reliability estimates were highly accurate at $r(307) = .89$, 95% CI [.84, .94] (manifest $r = .89$ [.86;.91]). Again, the SBERT model had substantially lower accuracy (manifest $r = .38$ [.28;.48]). Accuracy was lower when we excluded the randomly formed scales (manifest $r = .63$ [.50;.73]), as expected owing to the restricted range in the real scales ($SD = .10$ compared to $SD = .23$ in the combined set).

We subsequently investigated the model's validity for scale-level predictions using the holdout dataset. We averaged the vector representations of all items in each scale and then computed the cosine similarity of these averaged vectors. The convergence between empirical and synthetic scale correlations was remarkably high, exhibiting an accuracy of $r(6,245) = .89$ [.88, .90] (manifest correlation $r = .87$ [.86;.87]). 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 (*i.e., adopting the notion of distributional semantics which considers all contextual patterns as inherently semantic*). Again, the SBERT model had substantially lower accuracy (manifest $r = .33$ [.30;.35]).

In summary, the LLM-based synthetic estimates closely approximated the empirical inter-item and inter-scale correlations as well as reliability estimates and were robust to the checks detailed in Supplementary Note 2. Comparing predictions between the datasets used

in this pilot study leads us to expect that the effects are robust and will generalise to new, previously unseen English-language items.

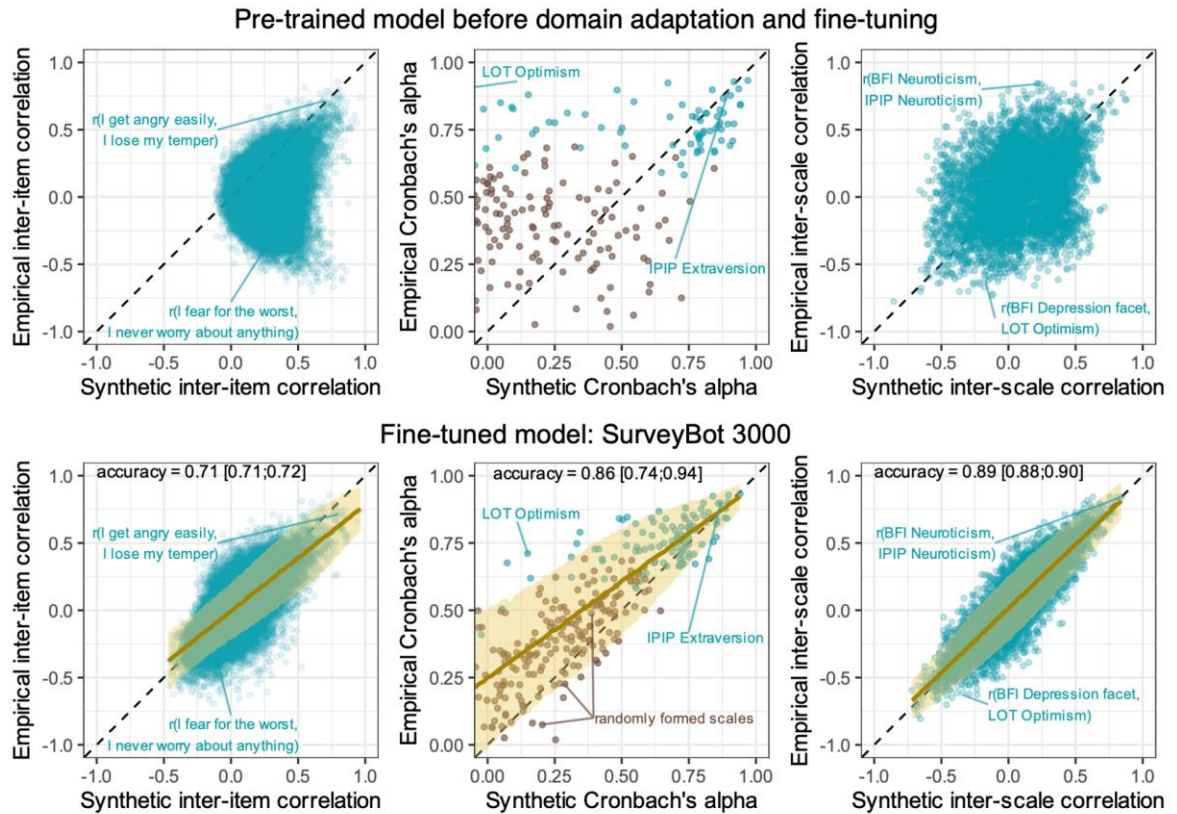


Figure 2. Scatter plots of the synthetic and empirical estimates, pilot study (Stage 1).

We show $N=87,153$ item pair correlations, $N=307$ scale reliabilities, and $N=6,245$ scale pair correlations for the pre-trained SBERT model (first row) and the fine-tuned SurveyBot3000 model (second row). The yellow line and shaded yellow region show the prediction and the 95% prediction interval for the latent outcome according to a Bayesian multi-membership regression model that allowed for heteroskedasticity and sampling error. Because the empirical estimates are estimated with sampling error, which the model adjusts for, fewer than 95% of dots are in the shaded prediction interval. Brown dots in the middle column show randomly combined scales, which we used to increase variance in the criterion. For reliabilities, 18 randomly combined scales with negative synthetic alphas according to the pre-trained model are not shown for ease of presentation.

Design

The primary objective of our research **was** to test the generalisability of our model in predicting human response patterns within survey data, that is, empirical item and scale correlations, as well as scale reliabilities. Our model's initial training data and our holdout represent a limited subset of the broader universe of survey items, with a skew towards personality psychology. We designed our validation study to challenge the model's capabilities by sampling from a more varied array of psychological measures. We **have collected** empirical data from a large online sample of English-speaking US Americans, similar to most of the studies in our training data. Participants **processed** the scales in random order, with item order randomised in each scale. While we **anticipated** a modest reduction in effect size during Stage 2 compared to the outcomes observed in the pilot study, we **expected** that the LLM-based synthetic estimates **would** still be sufficiently accurate to be useful. We present a Design Table summarising our methods and benchmarks.

Measures

To identify appropriate measures for our study, we conducted a comprehensive search of the APA PsycTests database. Our inclusion criteria for selecting scales were: a) utilisation of rating scales as the response format, b) items composed in the English language, c) scales developed within the last 30 years to minimise confounding factors related to changes in the English language, d) measures applicable to the general population, thus excluding scales only applicable to narrow target demographics such as adoptive parents or particular professional groups, e) measures applicable to a broad domain, thus excluding scales designed to rate specific consumer products or specific social attitudes, and f) freely accessible, non-proprietary measures. These criteria were mainly intended to make it feasible to have an unselected sample respond to most items. Within these constraints, we sampled scales to cover a wide range of measures used in the social and behavioural sciences.

We did not always use all items in a scale, so that we would be able to have participants respond to a large number in a scale. We included measures from industrial/organisational psychology, such as the Utrecht Work Engagement scale, measures from social psychology such as the Moral Foundations Questionnaire, from developmental psychology, such as the Revised Adult Attachment Scale, from clinical psychology, such as the Center for Epidemiological Studies Depression Scale, from emotion psychology, such as the positive and negative affect schedule, from personality psychology, such as Honesty-Humility in the HEXACO-60, and from other social sciences, such as the Attitudes Toward AI in Defence Scale and the Survey Attitude Scale. A full list of all scales can be found in Supplementary Note 5 and all items were deposited on OSF. In all, we **aimed** to have participants answer 246 items distributed across 79 scales and subscales.

Where possible, we adapted the response format to a 6-point Likert scale from *strongly disagree* to *strongly agree*. For the PANAS, CES-D, and the PSS, we used a 6-point scale from “never” to “most of the time” to better fit the item content. Our guiding principle was that a more uniform presentation was more important than a perfectly faithful rendering of the original scale. In addition, our current model is unaware of differing response formats and cannot account for them.

Sampling Plan

We used simulations to determine our number of scales, items, and survey participants. We wanted to precisely estimate the accuracy with which our synthetic estimates could approximate empirical estimates of inter-item and inter-scale correlations. Sampling error at the participant level affects the standard error with which we estimate empirical inter-item and inter-scale correlations and therefore would bias our accuracy estimates downward. To estimate empirical individual item correlations, we planned to use an online panel provider to collect a US quota sample of $N = 450$, before exclusions. In a quota sample, the panel provider attempts to approximately match the sample proportions to population proportions on three demographic variables: age, sex, and ethnicity. We had planned to limit participant recruitment to participants who have an approval rate exceeding 99% and have participated in at least 20 previous studies according to the sample provider, *Prolific*. However, this screener could not be combined with a quota sample, so no such limits were applied during recruitment. We paid participants regardless of whether they failed attention checks or completed the survey too quickly. In our planned analyses, we then estimated the accuracy of our manifest synthetic estimates for latent, error-free empirical estimates (see Supplementary Note 1).

From the APA PsycTests corpus, we sampled 246 items, which can be aggregated to 56 scales consisting of at least three items. We assumed we would retain a sample of at least $n = 400$ after exclusions. With the resulting 30,135 unique item pairs, we **expected** to infer the accuracy of our synthetic inter-item correlations to a precision (standard error) of ± 0.004 , according to our simulations. **Supplementing** our 57 scales with 200 randomly constituted scales, **enabled us to** infer the accuracy of our synthetic reliability estimates to a precision of ± 0.03 . With the resulting 1,568 unique scale pairs, without scale-subscale pairs, we **aimed to** infer the accuracy of our synthetic inter-scale correlations to a precision of ± 0.007 . The achieved precision is sufficient to detect even subtle deterioration in accuracy compared to our pilot study estimates.

Analysis Plan

We followed recommendations by Goldammer et al. (2020) and Yentes (2020) for identifying and excluding participants exhibiting problematic response patterns (e.g., careless responding). Accordingly, participants **were** excluded if any of the following conditions **were** met: a) participants voluntarily indicated that they did not respond seriously, b) multivariate outlier statistic using Mahalanobis distance, exceeding a threshold set for 99%

specificity), c) psychometric synonyms (defined as item pairs with $r > .60$) correlate below $r = .22$ for the participant), d) psychometric antonyms (defined as item pairs with $r \leq -.40$) correlate above $r = -.03$, e) low personal even-odd-index across scales ($r \leq .45$) f) average response times below 2 seconds per item. We checked the robustness of our conclusions to differently defined exclusion criteria.

We then computed all empirical inter-item correlations, inter-scale correlations, and reliabilities. Inter-item correlations used Pearson's product-moment correlations. We aggregated scales as the means of their items after reversing reverse-coded items. Inter-scale correlations were then computed as manifest Pearson's product-moment correlations. Reliability was estimated with the Cronbach's alpha coefficient based on inter-item correlation. We have uploaded synthetic estimates of the SBERT model and the SurveyBot3000 model for all of these coefficients to the OSF. The code for our preregistered analyses mirrored the code from our pilot study, including the robustness checks detailed in Supplementary Note 2. We planned to freeze both code and point predictions as part of our preregistration, but owing to a miscommunication between the two co-authors, nobody froze the repository and only point predictions for item correlations were uploaded to OSF. Because we discovered typographical errors in our version of the Moral Foundation Questionnaire, we revised the related point predictions after Stage 1 acceptance. After data collection, we merged empirical and synthetic estimates.

The central performance metric in this study is accuracy, defined as the convergence between synthetic and empirical estimates (not to be conflated with evaluation metrics of binary classifiers). We thus refer to *manifest accuracy* as the Pearson correlation between synthetic and empirical coefficients. We quantified *latent accuracy* using two complementary approaches that account for sampling error in empirical estimates. First, we used a structural equation modeling (SEM) approach where we fixed the residual variance of empirical estimates to the average sampling error variance and allowed manifest synthetic estimates to correlate with the latent variable. Second, we disattenuated for the standard error of the empirical estimates using a Bayesian errors-in-variables model, which allows for heteroskedastic accuracy (see Supplementary Note 1). We used the latter model as our primary estimate for latent accuracy. We also report the prediction error for all three quantities, as well as a plot similar to Figure 2. We furthermore report manifest and latent accuracies of the SBERT model, which we used as a benchmark (see Design Table).

Table 1. Design Table

Question	Hypothesis	Sampling plan	Analysis Plan	Interpretation given to different outcomes
How accurate are LLM-based synthetic inter-item correlations?	The synthetic estimates will exhibit an accuracy of $r = .71$ for the empirical inter-item correlation coefficients obtained from survey data, as estimated in our Bayesian multi-membership regression model.	246 items. With the resulting 30,135 unique item pairs, we should be able to estimate accuracy with a precision of ± 0.004 . A quota sample of $N=400$ will be drawn to estimate empirical correlations.	A correlation between synthetic and empirical estimates, disattenuated for the sampling error in the empirical estimates.	<p>If the accuracy matches (i.e. ± 0.02) that found in our pilot study, this is evidence that the model generalises well to novel survey items, including those outside personality psychology.</p> <p>In the unlikely case that the accuracy exceeds that found in our pilot study, we would carefully discuss why, including the potential that crowdworkers use LLMs to respond.</p>
How accurate are LLM-based synthetic reliability coefficients (for scales consisting of at least three items)?	The synthetic estimates will exhibit an accuracy of $r = .86$ for the empirical Cronbach's alpha coefficients obtained from survey data, as estimated in our Bayesian regression model.	As above. With the available 57 scales, supplemented by 200 randomly formed scales, we should be able to estimate accuracy with a precision of ± 0.02 .		<p>If the accuracy deteriorates to within 60% of the r in the pilot, the model may still be useful but should be applied with caution when item content is unlike the training data. We will examine and discuss performance across subfields to understand the deterioration. Retraining the model on a broader corpus would be indicated for future research.</p>
How accurate are LLM-based synthetic inter-scale correlations (for scales consisting of at least three items)?	The synthetic estimates will exhibit an accuracy of $r = .89$ for the empirical inter-scale correlation coefficients obtained from survey data, as estimated in our Bayesian multi-membership regression model.	As above. With the resulting 1,558 scale pairs, we should be able to estimate accuracy with a precision of ± 0.007 .		<p>If the accuracy deteriorates to below 60% of the r in the pilot, our model does not generalise well. Retraining with a broader corpus would be needed before recommending the model for wider use.</p> <p>If the accuracy of our model is reduced below the accuracy of the pre-trained model, our model training procedure overfit despite our precautions. The model should not be recommended for practical use and we would reinvestigate our precautions.</p>

Note. We determined the planned precision to detect any deterioration in performance greater than .01 for item pair correlations. Because increasing the number of scales is costlier than increasing the number of items, the sensitivity for the reliability coefficients is a compromise with feasibility.

Results

We collected data from N=470 participants using Prolific's online participant recruitment system. Because a bug in our questionnaire disrupted participation for an initial batch of participants who later returned to the study, we exceeded our planned sample size of 450 (see Supplementary Note 7 on deviations from preregistration).

We preregistered overly strict exclusion criteria because we misread Goldammer et al. (2020). After applying the preregistered criteria, only n=136 participants remained. Therefore, we used an adapted set of criteria that more closely followed Goldammer et al.'s (2020) recommendations for our main analyses, so that n=387 remained (see Table S7 in supplemental section). However, results for item pair correlations were robust to different exclusion criteria definitions as well as including all participants (see Supplementary Note 9). After applying the adapted exclusion criteria, the remaining participants had a mean age of 46.96 (SD = 15.58, range 18-86) and were 47% male. Most (63%) participants identified as non-Hispanic White, 13% as Black, and 12% as Hispanic. Four participants reported no high school education, 46 had a high school degree, 80 had some college experience, and 257 reported three or more years of college experience. Further and more detailed demographic information can be found in the online codebook.

All participants responded to a set of 219 items. Twenty-eight percent of the sample (n=110) were unemployed (or students etc.). Participants who reported being employed answered an additional set of 27 items specific to employment. We calculated pairwise Pearson's product-moment correlations between all item pairs in this set. We tested the accuracy of the preregistered SurveyBot3000 synthetic correlations against the empirical correlations of the resulting 30,135 item pairs.

Item pair correlations: Adjusted for sampling error in the empirical data (see Supplementary Note 1), the model's synthetic correlations predicted the empirical inter-item correlations with an accuracy of $r = .59$ (95% CI [.58;.60], manifest correlation $r = .57$ [.56;.58], Figure 3). Accuracy deteriorated compared to the holdout in our pilot study (to 83% of the $r = .71$ in the pilot), but our model was still able to generalise to this diverse set of items. Figure 3 shows the prediction of item correlations through semantic similarity, as estimated by the SBERT and SurveyBot3000 models. The SBERT model had substantially lower accuracy in predicting inter-item correlations (accuracy of $= .33$ [.32;.34]). We also computed the prediction error of the SurveyBot3000 in our model, i.e. how far off predictions were after accounting for sampling error in the empirical correlations in our model. The average root mean square error (RMSE) was $.17$ [.17;.17]. However, prediction error was larger when synthetic correlations were middling (.00 to .60) and smaller when they were negative or larger than .60, see Figure 4.

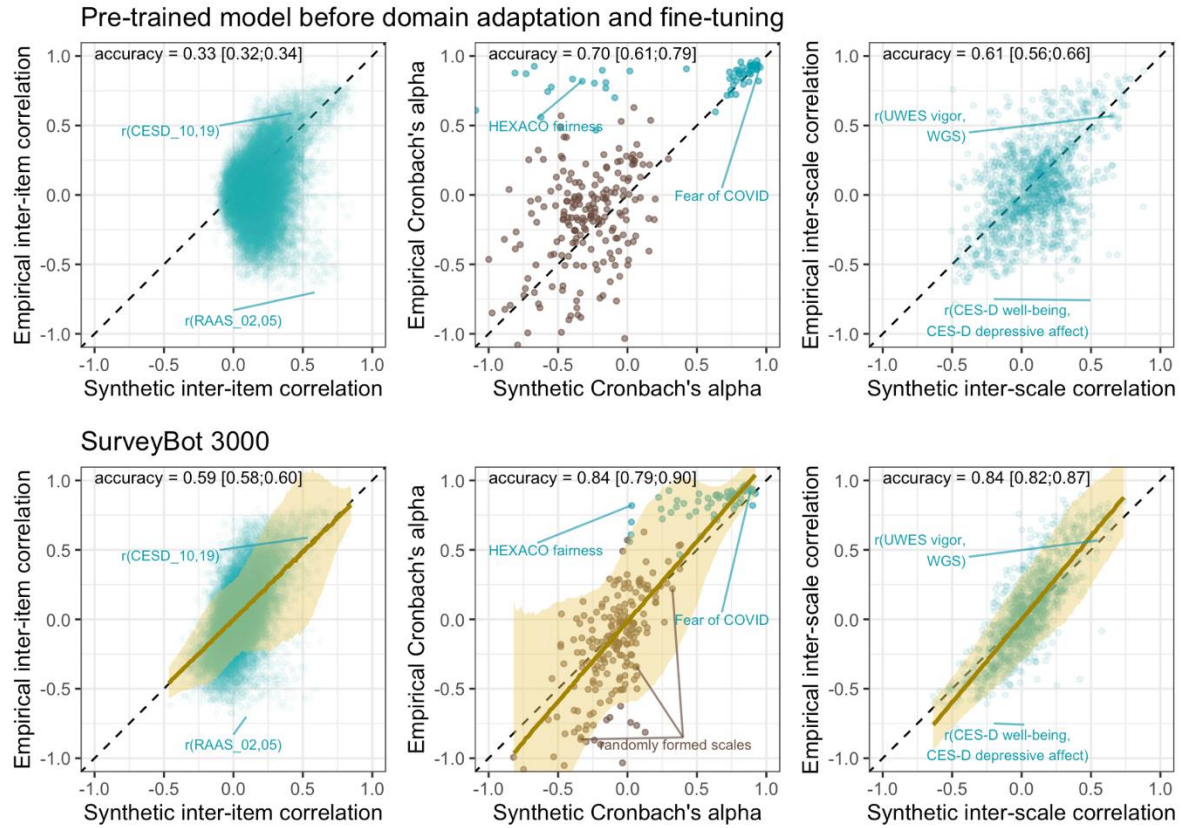


Figure 3. Scatter plots of the synthetic and empirical estimates, validation study (Stage 2). Showing $N=30,135$ item pair correlations, $N=257$ scale reliabilities, and $N=1,568$ scale pair correlations for the pre-trained SBERT model (first row) and the fine-tuned SurveyBot3000 model (second row).

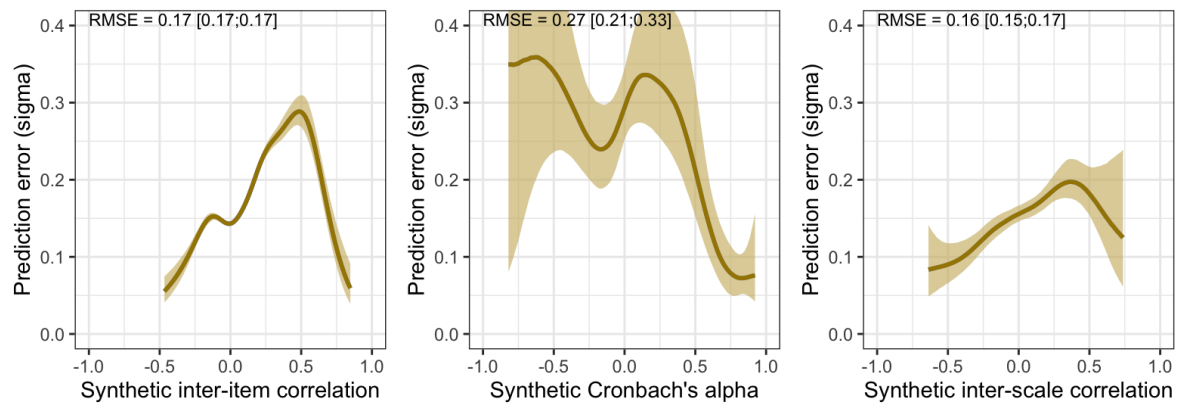


Figure 4. Prediction error of the synthetic estimates, validation study (Stage 2). Our prediction model allowed the error term to vary freely according to the predictor, the synthetic estimate. The thin-plate splines show that some synthetic estimates were predictably more accurate.

Scale reliabilities: We investigated the model's ability to predict scale reliabilities (Cronbach's alpha), which can be calculated from inter-item correlation matrices. For the 57 scales at least three, the manifest accuracy of the synthetic alpha coefficients was .64 [.45;.77]. This accuracy was slightly reduced compared to the pilot (94% of $r = .68$). Because all scales from the literature had restricted variability in reliability coefficients, we randomly sampled items to create 200 additional, varied scales. Unlike in the pilot, we reversed items randomly (not according to empirical correlations) and did not omit scales whose empirical Cronbach's alpha estimate was negative (see Table S7). We chose to make these changes to clarify that the synthetic alphas are in fact unbiased when we do not select on positive empirical alphas. We found that synthetic reliability estimates were highly accurate at $r(257) = .84$, 95% CI [.79, .90] (manifest $r = .85$ [.81;.88]). The SBERT model had lower accuracy than the SurveyBot3000 but performed much better than in the pilot study (manifest $r = .64$ [.56;.71]). The average root mean square error of the SurveyBot3000 estimates (RMSE) was .27 [.21;.33]. However, prediction error dropped below .10 when synthetic alphas entered the range seen in the real scales (above .60).

Scale pair correlations: We investigated the model's validity for scale-level predictions. For all scales with at least three items, we averaged the vector representations of all items (after reversing reverse-scored items) and then computed the cosine similarity of these averaged vectors. The accuracy of synthetic scale correlations was $r(1,568) = .84$ [.82, .87] (excluding scale-subscale pairs; manifest correlation $r = .83$ [.81, .85]) our fine-tuned LLM explained 71% of the latent variance in scale intercorrelations, based on nothing but semantic information contained in the items. Manifest accuracy for the 228 scale pairs where each scale had at least five items was $r = .88$). Performance was slightly attenuated compared to the pilot (94% of $r = .89$), but this may be partly because scales in this set were slightly shorter (mean number of items = 5.75) than in the pilot (6.79), see also Supplementary Note 8. As for synthetic reliabilities, the SBERT model had lower accuracy than the SurveyBot3000 but performed much better than in the pilot study (manifest $r = .50$ [.46;.54]). The average root mean square error of the SurveyBot3000 estimates (RMSE) was .16 [.15;.17]. As for item correlations, prediction error was larger for middling synthetic estimates (.00 to .50) than for negative and high positive estimates (Figure 4).

By domain

We investigated the accuracy of our synthetic inter-item correlations by domain. We had grouped scales into five domains (attitudes, personality, clinical, social, and occupational psychology). Manifest accuracy was lowest for attitudes ($r = .34$ within the attitude domain, $r = .31$ when attitude items were correlated with items in other domains) and highest for occupational psychology ($r = .75$ within, $r = .65$ across). In all domains, the SurveyBot3000 predictions outperformed the SBERT predictions, so there was no obvious trade-off between fine-tuning and generalisability (see Figure 5).

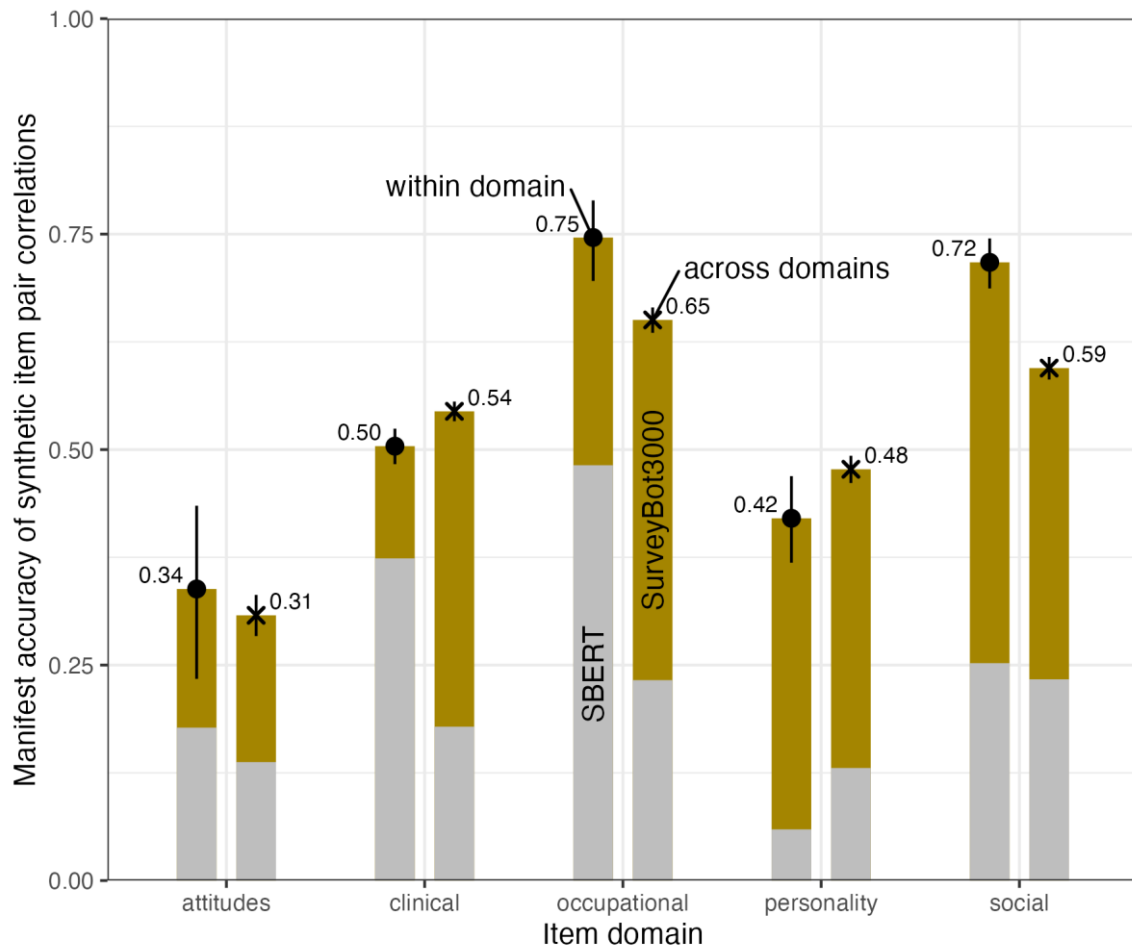


Figure 5: Accuracy by domain. Accuracy differed across domains. SurveyBot3000 accuracy (colored) was always higher than SBERT accuracy (gray). Results were largely consistent whether accuracy of items was tested within domains (left, circle) or across domains (right, cross).

Robustness checks

We repeated all robustness checks we conducted for the pilot study and added additional checks. Because we had preregistered overly strict exclusion criteria and as we were unable to combine quota sampling with a screener for highly rated Prolific participants, we estimated the accuracy of the synthetic item correlations after applying different sets of defensible exclusion criteria. After accounting for sampling error, accuracy varied between .57 and .59 depending on the exclusion criteria, i.e. not substantially (Figure 6, see also Supplementary Note 9). We report further robustness checks and sensitivity analyses in Supplementary Note 8.

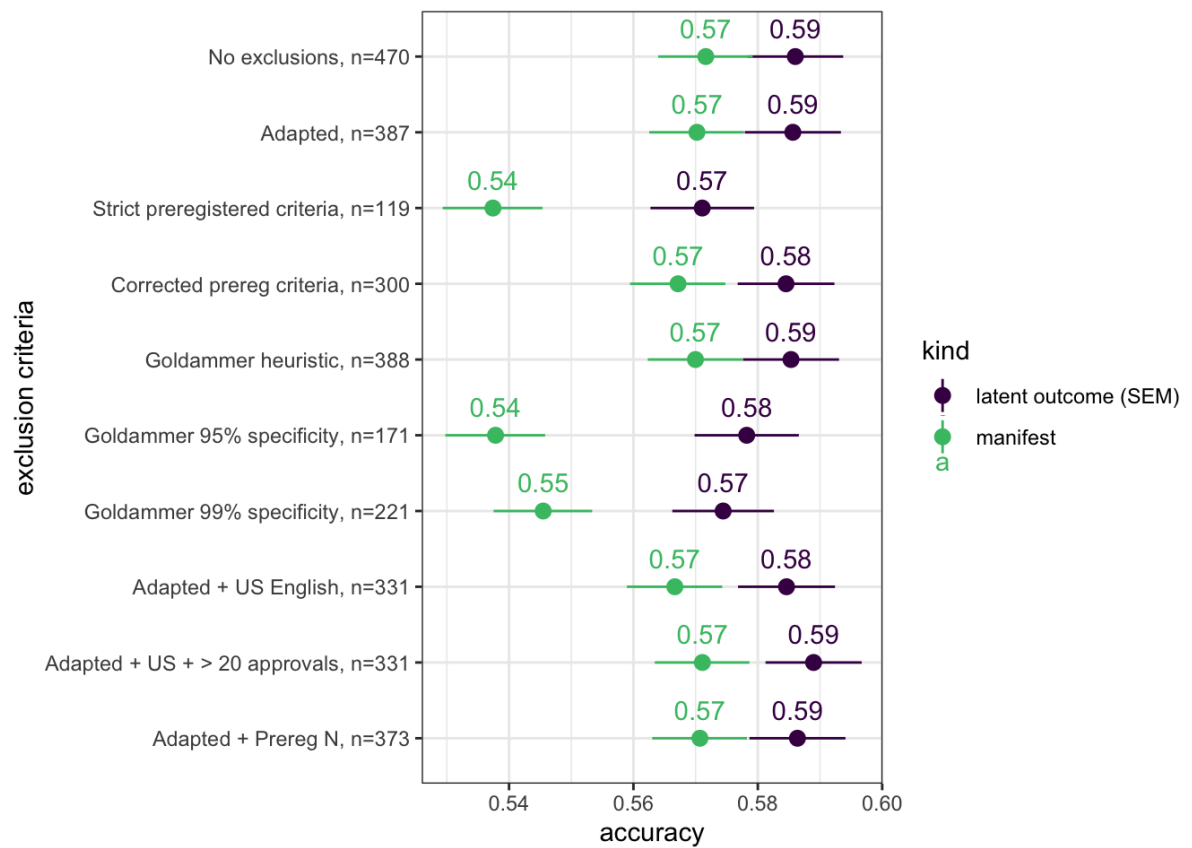


Figure 6. Applying different exclusion criteria (or none) did not cause large changes in the estimated latent accuracy (see Supplementary Note 7). Predictably, manifest accuracy was reduced when we excluded many participants.

Discussion

We introduce a computational linguistics approach that synthetically predicts associations between survey responses—including item-level correlations, scale-level relationships, and derived psychometric properties—with high accuracy. Using our SurveyBot3000, these synthetic estimates have a margin of error that is comparable to a small pilot study, but free and instant. Our preregistered validation study confirms the convergence between synthetic predictions and empirical datasets, validating the method's ability to mirror real-world reliability coefficients, scale correlations, and covariance patterns, even outside the content domain of personality psychology.

Accuracy in our preregistered validation was attenuated compared to our pilot study (up to 83% of the pilot study's accuracy for item pairs) but never to the level of the pre-trained model. So, even though the items spanned a broader domain, the synthetic estimates had margins of error comparable to a small pilot study. Attenuation was strongest for item pairs ($r = .59$ [.58;.60]). After aggregation, accuracy was higher for scale pairs (latent $r = .84$ [.82;.87]) and for reliabilities ($r = .84$ [.79;.90]; attenuation to 94% of the pilot study's accuracy). Our prediction model allowed for the margin of error to depend on the synthetic estimate. Indeed, because the SurveyBot3000 still sometimes predicts positive correlations instead of negative correlations, negative synthetic estimates are more accurate (see Figure 4). For instance, a negative synthetic scale correlation is estimated about as accurately as in a $N=80$ pilot study, whereas a positive correlation is only about as accurately estimated as a $N=20$ pilot study (see Supplementary Note 10). The margin of error was also larger for synthetic reliabilities below commonly used cutoffs, i.e. $< .60$).

Recent related contributions on computational modeling for survey research (e.g. Hernandez & Nie, 2023; Schoenegger et al., 2024; Wulff & Mata, 2023), highlight the field's growing interest in synthetic prediction of psychometric patterns. In a recent update to their work, Wulff & Mata (2025) have adopted fine-tuning techniques that improve upon their earlier results, yielding accuracies that approach the performance we report here, but limited to absolute correlations. In another parallel effort, Schoenegger et al. (2025) report comparable performance of the proprietary model PersonalityMap and the SurveyBot3000. However, this comparison is difficult to interpret because the SurveyBot3000 was trained on the data used as the test set and the PersonalityMap model is proprietary, which makes it difficult to assess leakage and generalizability.

Our work advances this area of synthetic survey modeling not mainly by reporting top-tier accuracy but through methodological innovations and practical tools designed to improve the rigor, transparency, generalizability, and accessibility.

First, we introduce a two-step training protocol that refines sentence transformer models for robust prediction of survey response associations. Key safeguards include training on a diverse item corpus to minimize domain bias, strict contamination controls to prevent overfitting, and systematic hyperparameter optimization. A novel calibration step further enables the model to predict negative correlations (e.g., opposing items), more accurately reflecting the empirical distribution of coefficients. The resulting model, the SurveyBot3000,

demonstrates performance exceeding known human capabilities in correlation judgment (Epstein & Teraspulsky, 1986).

Second, to ensure transparency and minimize analytic flexibility, we preregistered our validation protocol and underwent formal Stage 1 peer review prior to testing. This safeguards against overfitting and confirms that accuracy claims are not artifacts of post hoc adjustments.

Third, we systematically evaluate generalizability across psychological domains, including personality, clinical, and social psychology, as well as social attitudes. While item-level accuracy varies with conceptual diversity—attenuated in cross-domain tests compared to our pilot study—the SurveyBot3000 always outperformed the pre-trained baseline model (i.e. SBERT), so our fine-tuning did not impede generalisability.

Finally, we provide an open-access web application (<https://huggingface.co/spaces/magnolia-psychometrics/synthetic-correlations>) to democratize access to synthetic psychometric predictions. The tool generates immediate estimates of internal consistency, scale structure, and inter-item correlations from text inputs, offering researchers a free pretesting resource with guidance for responsible interpretation. The application can be considered a free pilot study of survey items to investigate factor structure and internal consistency. Similar to pilot studies, synthetic estimates can tell us “where to look” but should always be followed up with more empirical data before conclusions are drawn.

As the behavioural sciences grapple with an ever-expanding universe of oftentimes redundant measures, our line of research has the potential to re-organise the vast collection of scales accumulated over the past decades of research and to help prevent further proliferation and fragmentation in the future (Elson et al., 2023; Anvari et al., 2024; Anvari et al., in press). Rosenbusch et al. (2020) laid important groundwork on computational language-based methods to semantically search for psychometric scales, but were constrained by the technological limitations of their time. Our results and work on the SurveyBot3000 encourages us that the technological foundation for such an ambitious undertaking has matured.

The APA PsycTests database currently holds over 78,000 records, with the majority of scales only being used once or twice (Elson et al., 2023; Anvari et al., 2024). With both the methodology and the data in place, we propose that future research efforts should be dedicated towards the development of a semantic search engine. Searching such a “synthetic nomological net” could reveal potential overlap between tens of thousands of items and scales and ultimately help us avoid redundancy and confusing labels. A more parsimonious ontology could then enable better evidence synthesis. A semantic search engine could be a tool in the scale development and the peer-review process, in order to help authors and reviewers to assess the incremental value of newly developed scales and proposed constructs. Potential redundancies and confusing labels (e.g., jingle/jangle

fallacies; Wulff & Mata, 2023) could then be flagged for empirical follow-up. Importantly, such a system would make the search problem tractable. That is, the SurveyBot3000 could help pick scales out of the ten thousands in existence to empirically evaluate the novel scale for discriminant validity. That way, humans remain in the loop. We believe that this line of work exemplifies a responsible integration of LLMs into research, which is a topic of current debate (Binz et al., 2023). Specifically, the collaborative circumstances in scale development carry minimal risk for harmful effects on the scientific ecosystem. False negatives (i.e., the model fails to detect redundant scales) would merely maintain the status quo, which has led to construct proliferation in the first place. False positives (e.g., the model incorrectly flags two measures as redundant) would require researchers to verify this empirically before drawing conclusions. This balanced approach, where LLMs accelerates discovery while human researchers retain interpretive authority, should characterize a productive human-AI collaboration across the social and behavioural sciences.

To further strengthen the potential of computer linguistic approaches to survey pattern prediction, we noted some limitations in the SurveyBot3000 that need to be addressed by future research. Despite the strong convergence between synthetic and empirical data in both the pilot and validation study, the SurveyBot3000 occasionally struggled to infer negative correlations.

While polarity calibration clearly improved the model's handling of negatively worded items overall (see Supplementary Note 6), the synthetic estimates still had a bias towards positive signs. Of the empirical correlations, 59% were positive, whereas 67% of the synthetic correlations were. In keeping with this, a negative synthetic item correlation predicted the empirical sign incorrectly slightly less often (16%) than a positive synthetic item correlation (19%). If we imagine that a human user of our app can correct the coefficient sign in these small-scale applications, this would improve manifest accuracy by .11, yielding an overall convergence of .68 between synthetic estimates and empirical correlations.

Various linguistic aspects were associated with impaired predictions, but no clear pattern emerged. For example, items that avoided self-directed language were predicted less accurately. However, for such items, we did not observe any increase in accuracy after rephrasing the statements to use first-person pronouns (see Supplementary Note 8). In the current study, item length, self-directedness, sentence complexity and content domain are all confounded with one another. Further efforts could be directed towards systematically manipulating and investigating lexicographic (e.g., grammatical form, item length) and item-metric (e.g., observability, temporality; Leistner et al., 2024; Leising et al., 2014) features potentially influencing accuracy in survey pattern prediction independently of content domain (Hommel, 2024).

Both the sign prediction errors and accuracy fluctuations arising from unconventional linguistic aspects could potentially be addressed by recent innovations. For example, Opitz & Frank (2022) have shown that vector representations of text can be decomposed into explainable semantic features. Instead of comparing vectors monolithically, future approaches could isolate psychometrically relevant information by separating residual features in vector space. This decomposition approach may help establish theoretical upper

bounds on prediction accuracy by distinguishing between different types of semantic content captured in vector space, including conceptual meaning, but also peripheral semantic information such as survey response tendencies.

Beyond these technical refinements, model performance could be enhanced through a more balanced training corpus, as suggested by domain-specific variations in predictive accuracy. For instance, synthetic estimates for clinical psychology measures performed worse than for social psychology measures, reflecting the limited representation of psychopathology items in our training data. Balancing the corpus aligns with established principles of language model development where capabilities consistently improve with increased training data, model size, and computational resources (Kaplan et al., 2020). However, the same note of caution as above applies, because content domain is confounded with lexicographic and item-metric aspects. In addition, the low accuracy of synthetic estimates in the attitude domain can be partly attributed to the fact that attitude items have lower absolute intercorrelations on average, so there is less variance to explain. On the root mean square error metric of accuracy, which does not have this issue, attitude items had middling accuracy compared to other domains.

Robust evaluation protocols are essential to systematically assess and compare the capabilities and limitations of current and future model developments. To this end, benchmark tests are usually established for specific tasks related to language modelling using infrastructure providers like Hugging Face (*Hugging Face Datasets*, n.d.) and Kaggle (*Kaggle Datasets*, n.d.). We recommend that efforts should be undertaken to develop such a standardized holdout set to objectively track future progress in survey pattern prediction with comparable accuracy metrics. Although many currently available fine-tuned models are trained on the same or overlapping data (chiefly SAPA, Condon, Roney, & Revelle, 2017), it is currently difficult to compare models because teams divide training and test partitions differently, i.e. one model is trained on the data that another team uses as its benchmark. For fair comparisons, we need transparency about the contents of the training data, including for proprietary models, or ways to come up with guaranteed novel items.

Our final report deviates from our pre-planned Stage 1 protocol in several ways. We transparently communicated these deviations according to Willroth and Atherton (2024) in Supplementary Note 7 and reported additional robustness checks to study the impact of these deviations on our results. We found that latent accuracy was largely unaffected after readjusting exclusion criteria and generally conclude that the deviations had little impact.

Sentence transformers can effectively model psychometric properties and relationships using solely the semantic information contained within item texts. Our work establishes a method that produces synthetic predictions which converge with empirical survey data and demonstrate robust generalization beyond the training domain. We see many potential applications, simplified through the web app we have released. The SurveyBot3000's synthetic estimates have a margin of error comparable to a small pilot study. As with pilot studies, the synthetic estimates can guide an investigation but need to be followed up by human researchers with human data. By making synthetic estimates freely available, we hope to reduce ad hoc measurement culture. Researchers should now find it easier to

compare existing measures, and to identify old and new measures with desirable psychometric properties.

Looking ahead, incorporating recent advances in computational linguistics may yield increasingly precise models that could serve as foundational tools for untangling the nomological net (Cronbach & Meehl, 1955) and constructing a unified taxonomy of psychological measures.

Data availability

We have shared all key materials on the Open Science Framework at <https://osf.io/z47qs/>. The existing data used for training and in the pilot study has been openly shared, we link to the original sources. Anonymized data for the validation study have also been shared on OSF.

Code availability

We have shared the training and analysis code on the Open Science Framework at <https://osf.io/z47qs/>.

Acknowledgements

The research is funded by the German Research Foundation grant #464488178 to Ruben C. Arslan. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank Stefan Schmukle, Anne Scheel, Julia Rohrer, Malte Elson, Taym Alsalti, Ian Hussey, Saloni Dattani, David Condon, Dirk Wulff and Jan Arnulf for helpful discussions. We also thank Jan-Paul Ries, Lorenz Oehler, and Sarah Lennartz for comments on an earlier version of this manuscript.

Author contributions

B.E.H.: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Visualization, Writing - original draft, and Writing - review & editing.

R.C.A.: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Validation, Visualization, Writing - original draft, and Writing - review & editing.

Competing interests

Björn E. Hommel is affiliated with magnolia psychometrics GmbH, a private consulting agency which has agreed to maintain the app accompanying this paper, currently hosted on Hugging Face. There are no competing interests, financial or otherwise, related to this research. The SurveyBot3000 model is freely licensed under Apache 2.0. Potential future commercial applications of these findings may be developed.

References

- Abdurahman, S., Vu, H., Zou, W., Ungar, L., & Bhatia, S. (2024). A deep learning approach to personality assessment: Generalizing across items and expanding the reach of survey-based research. *Journal of Personality and Social Psychology*, 126(2), 312–331.
<https://doi.org/10.1037/pspp0000480>
- Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). *Optuna: A Next-generation Hyperparameter Optimization Framework* (No. arXiv:1907.10902). arXiv.
<https://doi.org/10.48550/arXiv.1907.10902>
- Anvari, F., Alsalti, T., Oehler, L., Hussey, I., Elson, M., & Arslan, R. C. (2024). *A fragmented field: Construct and measure proliferation in psychology*. <https://doi.org/10.31234/osf.io/b4muj>
- Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J., Rytting, C., & Wingate, D. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3), 337–351.
<https://doi.org/10.1017/pan.2023.2>
- Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L., & Bong, C. H. (2014). Predicting Survey Responses: How and Why Semantics Shape Survey Statistics on Organizational Behaviour. *PLoS ONE*, 9(9), e106361. <https://doi.org/10.1371/journal.pone.0106361>
- Bainbridge, T. F., Ludeke, S. G., & Smillie, L. D. (2022). Evaluating the Big Five as an organizing framework for commonly used psychological trait scales. *Journal of Personality and Social Psychology*, 122(4), 749–777. <https://doi.org/10.1037/pspp0000395>
- Berufsverband Deutscher Psychologinnen und Psychologen. (2022). *Ethische Richtlinien der Deutschen Gesellschaft für Psychologie e. V. und des Berufsverbandes Deutscher Psychologinnen und Psychologen e. V.*[Ethical guidelines of the German Society for Psychology and the Professional Association of German Psychologists]. DGPs.
<https://www.dgps.de/die-dgps/aufgaben-und-ziele/berufsethische-richtlinien/>
- 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. (2023). *How should the advent of large language models affect the practice of science?* (No. arXiv:2312.03759). arXiv.

<https://doi.org/10.48550/arXiv.2312.03759>

Condon, D. M. (2017). *The SAPA Personality Inventory: An empirically-derived, hierarchically-organized self-report personality assessment model*. PsyArXiv.

<https://doi.org/10.31234/osf.io/sc4p9>

Condon, D. M., & Revelle, W. (2015). Selected Personality Data from the SAPA-Project: On the Structure of Phrased Self-Report Items. *Journal of Open Psychology Data*, 3.

<https://doi.org/10.5334/jopd.al>

Condon, D. M., Roney, E., & Revelle, W. (2017). A SAPA Project Update: On the Structure of phrased Self-Report Personality Items. *Journal of Open Psychology Data*, 5(1), Article 1.

<https://doi.org/10.5334/jopd.32>

Credé, M., Tynan, M. C., & Harms, P. D. (2017). Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and Social Psychology*, 113(3), 492–511.

<https://doi.org/10.1037/pspp0000102>

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302. <https://doi.org/10.1037/h0040957>

Cutler, A., & Condon, D. M. (2022). Deep lexical hypothesis: Identifying personality structure in natural language. *Journal of Personality and Social Psychology*.

<https://doi.org/10.1037/pspp0000443>

Digman, J. M. (1990). Personality Structure: Emergence of the Five-Factor Model. *Annual Review of Psychology*, 41(1), 417–440. <https://doi.org/10.1146/annurev.ps.41.020190.002221>

Elson, M., Hussey, I., Alsalti, T., & Arslan, R. C. (2023). Psychological measures aren't toothbrushes. *Communications Psychology*, 1(1), Article 1. <https://doi.org/10.1038/s44271-023-00026-9>

- Epstein, S., & Teraspulskey, L. (1986). Perception of cross-situational consistency. *Journal of Personality and Social Psychology*, 50(6), 1152. <https://doi.org/10.1037/0022-3514.50.6.1152>
- Flake, J. K., & Fried, E. I. (2020). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science*, 3(4), 456–465. <https://doi.org/10.1177/2515245920952393>
- Fyffe, S., Lee, P., & Kaplan, S. (2024). “Transforming” Personality Scale Development: Illustrating the Potential of State-of-the-Art Natural Language Processing. *Organizational Research Methods*, 27(2), 265–300. <https://doi.org/10.1177/10944281231155771>
- Greene, R., Sanders, T., Weng, L., & Neelakantan, A. (2022). New and improved embedding model. *Open AI Blog*. <https://openai.com/blog/new-and-improved-embedding-model>
- Guenole, N., D’Urso, E. D., Samo, A., & Sun, T. (2024). *Pseudo Factor Analysis of Language Embedding Similarity Matrices: New Ways to Model Latent Constructs*. OSF. <https://doi.org/10.31234/osf.io/vf3se>
- Hernandez, I., & Nie, W. (2023). The AI-IP: Minimizing the guesswork of personality scale item development through artificial intelligence. *Personnel Psychology*, 76(4), 1011–1035. <https://doi.org/10.1111/peps.12543>
- Hommel, B. E. (2024). *The advent of transformer models in psychometrics* [PhD Thesis, lmu]. <https://edoc.ub.uni-muenchen.de/33252/>
- Hommel, B. E., Wollang, F.-J. M., Kotova, V., Zacher, H., & Schmukle, S. C. (2022). Transformer-Based Deep Neural Language Modeling for Construct-Specific Automatic Item Generation. *Psychometrika*, 87(2), 749–772. <https://doi.org/10.1007/s11336-021-09823-9>
- Hugging Face Datasets* (n.d.). (2025, January 24). <https://huggingface.co/docs/datasets/en/index>
- Hugging Face model hub*. (n.d.). Retrieved June 2, 2023, from <https://huggingface.co/sentence-transformers/all-mpnet-base-v2>

- Hussain, Z., Binz, M., Mata, R., & Wulff, D. U. (2023). *A tutorial on open-source large language models for behavioral science*. PsyArXiv. <https://osf.io/preprints/psyarxiv/f7stn>
- Hussey, I., Alsalti, T., Bosco, F., Elson, M., & Arslan, R. C. (2024). *An aberrant abundance of Cronbach's alpha values at .70*. <https://doi.org/10.31234/osf.io/dm8xn>
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling Laws for Neural Language Models*. <https://doi.org/10.48550/ARXIV.2001.08361>
- Kjell, O. N. E., Kjell, K., Garcia, D., & Sikström, S. (2019). Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs. *Psychological Methods*, 24(1), 92.
- Kjell, O. N. E., Kjell, K., & Schwartz, H. A. (2024). Beyond rating scales: With targeted evaluation, large language models are poised for psychological assessment. *Psychiatry Research*, 333, 115667. <https://doi.org/10.1016/j.psychres.2023.115667>
- Kopalle, P. K., & Lehmann, D. R. (1997). Alpha Inflation? The Impact of Eliminating Scale Items on Cronbach's Alpha. *Organizational Behavior and Human Decision Processes*, 70(3), 189–197. <https://doi.org/10.1006/obhd.1997.2702>
- Larsen, K. R., & Bong, C. H. (2016). A Tool for Addressing Construct Identity in Literature Reviews and Meta-Analyses. *MIS Quarterly*, 40(3), 529–551. <https://doi.org/10.25300/MISQ/2016/40.3.01>
- Leising, D., Scharloth, J., Lohse, O., & Wood, D. (2014). What Types of Terms Do People Use When Describing an Individual's Personality? *Psychological Science*, 25(9), 1787–1794. <https://doi.org/10.1177/0956797614541285>
- Leistner, M., Hommel, B. E., Wendt, L. P., & Leising, D. (2024). *Properties of Person Descriptors in the Natural German Language: A Preregistered Replication and Extension*. PsyArXiv. <https://doi.org/10.31234/osf.io/s8h7u>

NVIDIA, Vingelmann, P., & Fitzek, F. H. P. (2022). *CUDA, release: 11.7.1*.

<https://developer.nvidia.com/cuda-toolkit>

Opitz, J., & Frank, A. (2022). *SBERT studies Meaning Representations: Decomposing Sentence*

Embeddings into Explainable Semantic Features (No. arXiv:2206.07023). arXiv.

<https://doi.org/10.48550/arXiv.2206.07023>

Pretrained Models—Sentence-Transformers documentation. (n.d.). Retrieved March 5, 2024, from

https://www.sbert.net/docs/pretrained_models.html

R Core Team. (2023). *R: A Language and Environment for Statistical Computing* (Version 4.3.0)

[Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-*

Networks (No. arXiv:1908.10084). arXiv. <https://doi.org/10.48550/arXiv.1908.10084>

Rosenbusch, H., Wanders, F., & Pit, I. L. (2020). The Semantic Scale Network: An online tool to detect

semantic overlap of psychological scales and prevent scale redundancies. *Psychological*

Methods, 25(3), 380–392. <https://doi.org/10/gg5rn7>

Schoenegger, P., Greenberg, S., Grishin, A., Lewis, J., & Caviola, L. (2024). *Can AI Understand Human*

Personality? -- Comparing Human Experts and AI Systems at Predicting Personality

Correlations (No. arXiv:2406.08170). arXiv. <https://doi.org/10.48550/arXiv.2406.08170>

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition

and clustering. *Proceedings of the IEEE Conference on Computer Vision and Pattern*

Recognition, 815–823. <https://www.cv->

[foundation.org/openaccess/content_cvpr_2015/html/Schroff_FaceNet_A_Unified_2015_CV](https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Schroff_FaceNet_A_Unified_2015_CV)

[PR_paper.html](https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Schroff_FaceNet_A_Unified_2015_CV)

Sharp, C., Kaplan, R. M., & Strauman, T. J. (2023). The Use of Ontologies to Accelerate the Behavioral

Sciences: Promises and Challenges. *Current Directions in Psychological Science*, 32(5), 418–

426. <https://doi.org/10.1177/09637214231183917>

- Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020). *MPNet: Masked and Permuted Pre-training for Language Understanding* (No. arXiv:2004.09297). arXiv.
<https://doi.org/10.48550/arXiv.2004.09297>
- Tunstall, L., Werra, L. von, Wolf, T., & Géron, A. (2022). *Natural language processing with Transformers: Building language applications with Hugging Face* (First edition). O'Reilly.
- Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, \Lukasz, & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 5998–6008. <https://arxiv.org/abs/1706.03762>
- Williams, A., Nangia, N., & Bowman, S. R. (2018). *A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference* (No. arXiv:1704.05426). arXiv.
<https://doi.org/10.48550/arXiv.1704.05426>
- Willroth, E. C., & Atherton, O. E. (in press). Best Laid Plans: A Guide to Reporting Preregistration Deviations. *Advances in Methods and Practices in Psychological Science*.
<https://doi.org/10.31234/osf.io/dwx69>
- Wulff, D. U., & Mata, R. (2023). *Automated jingle–jangle detection: Using embeddings to tackle taxonomic incommensurability*. PsyArXiv. <https://doi.org/10.31234/osf.io/9h7aw>
- Wulff, D. U., & Mata, R. (2025). Semantic embeddings reveal and address taxonomic incommensurability in psychological measurement. *Nature Human Behaviour*, 1–11.
<https://doi.org/10.1038/s41562-024-02089-y>