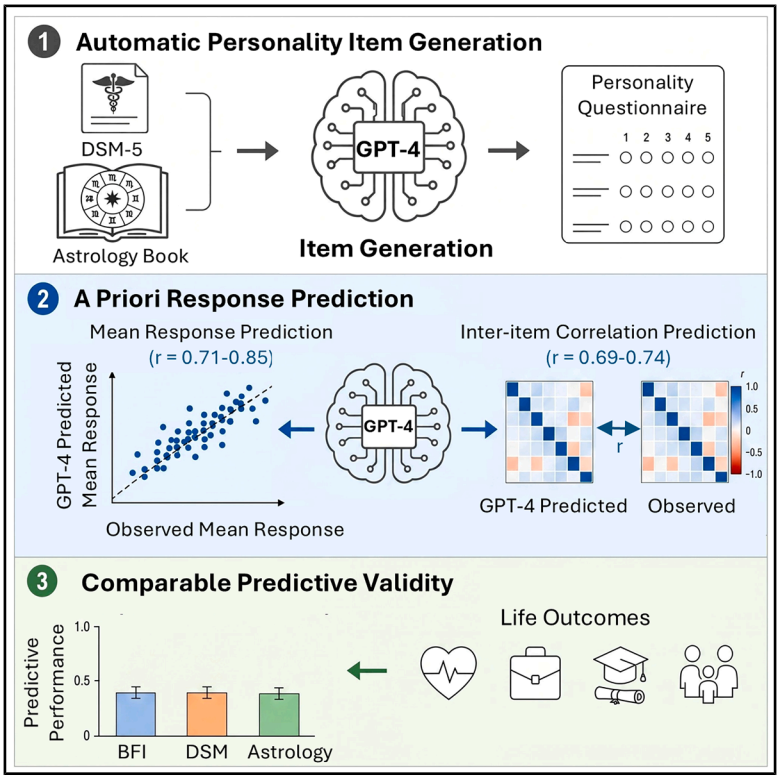
iScience

Article

Generating and analyzing personality questionnaires using large language models

Graphical abstract



Highlights

- We built an LLM tool to generate and validate personality questionnaires from text
- GPT-4 predicts mean responses and inter-item correlations before data collection
- DSM- and astrology-based items predict outcomes comparably to Big Five Inventory

Authors

Rotem Monsa, Aviv Zohar, Shahar Arzy

Correspondence

rotem.monsa@mail.huji.ac.il

In brief

Applied computing; Interdisciplinary application studies; Psychology

Monsa et al., 2026, iScience ■■, 116909
■■, 2026 © 2026 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.isci.2026.116909

Please cite this article in press as: Monsa et al., Generating and analyzing personality questionnaires using large language models, iScience (2026), https://doi.org/10.1016/j.isci.2026.116909

iScience

CellPress
OPEN ACCESS

Article

Generating and analyzing personality questionnaires using large language models

Rotem Monsa,[1,5,*] Aviv Zohar,[2] and Shahar Arzy[1,3,4]
[1]Computational Neuropsychiatry Lab, Department of Medical Neurosciences, Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem 9112001, Israel
[2]Department of Computer Science and Engineering, Hebrew University of Jerusalem, Jerusalem 9190401, Israel
[3]Department of Neurology, Hadassah Hebrew University Medical School, Jerusalem 9112001, Israel
[4]Department of Brain and Cognitive Sciences, Hebrew University of Jerusalem, Jerusalem 9190501, Israel
[5]Lead contact
*Correspondence: rotem.monsa@mail.huji.ac.il
https://doi.org/10.1016/j.isci.2026.116909

SUMMARY

The Five-Factor Model (“Big Five”) is based on the lexical hypothesis that personality traits are encoded in language. Large language models (LLMs) offer new ways to explore personality through text. We developed an LLM-based method to generate and validate personality questionnaires from textual sources. Using the DSM-5 personality disorders section and a popular astrology book, we generated two questionnaires and administered them, alongside the Big-Five Inventory (BFI), to 600 adults. Internal consistency was high for the BFI and the DSM-based questionnaire but low for the astrology-based one. The LLM predicted human response patterns for both generated questionnaires. Individual items from both the DSM- and astrology-based questionnaires predicted diverse life outcomes at levels comparable to the BFI. These findings show that LLMs can construct personality measures and anticipate response patterns, offering a scalable framework for corpus-based personality research.

INTRODUCTION

The Five-Factor Model (FFM) of personality, also known as the Big Five, is one of the most extensively validated and widely used frameworks for understanding human personality.[1,2] Its five dimensions - Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism - were solidified as robust dimensions of personality, and have demonstrated explanatory power across numerous studies and cultural contexts.[3] The Big-Five personality traits are rooted in the lexical hypothesis, first proposed by Sir Francis Galton in 1884, who suggested that important personality traits are encoded in natural language.[4] This idea was formalized by Allport and Odbert[5] who compiled a list of 4,504 adjectives describing personality traits. Further works gradually reduced the number of traits and applied factor analysis to identify major clusters,[6,7] finally reaching a five-factor structure[8,9] named the “Big Five.”[10] While the exact meaning of the five traits changed slightly between works,[9,11–13] conceptually the core traits of the FFM have remained consistent. Over time, the model has become widely accepted in modern psychology due to its cross-cultural stability and predictive validity across diverse behavioral domains.[14,15]

During the 2010s, personality psychology entered the computational era in which natural language processing (NLP) techniques (e.g., word embeddings, deep neural networks) were applied to large textual corpora and digital behavioral traces. Studies using social-media posts, blogs, or other web texts demonstrated that machine-learning models could predict Big Five traits from “digital footprints,” inaugurating a new paradigm of computational personality assessment.[16] Early approaches relied on NLP techniques such as word embeddings (e.g., Word2Vec, GloVe) and supervised models including shallow or deep neural networks to extract semantic patterns from text and relate them to personality traits.[17] While these methods provided proof-of-concept links between language and personality, their predictive performance and interpretability were often limited, partly due to shallow semantic representations and restricted ability to capture long-range context. These limitations motivated the use of large language models (LLMs), which are designed to capture richer semantic and contextual structure in language.[18]

The appearance of LLMs[19–21] trained on trillions of tokens representing the majority of human-generated textual content represents a leap forward in the ability to handle language-based tasks.[22–24] As such, LLMs embody an understanding of language, often matching that of human experts on various linguistic tasks. Given the lexical roots of the FFM, this raises the intriguing possibility that LLMs may capture statistical patterns reflecting human personality traits as a natural consequence of their training. More broadly, LLMs provide an unprecedented opportunity to model human personality directly from text by leveraging statistical regularities that correlate with human traits.[25]

Please cite this article in press as: Monsa et al., Generating and analyzing personality questionnaires using large language models, iScience (2026), https://doi.org/10.1016/j.isci.2026.116909

Using LLMs for personality questionnaire development requires that the generated items meet established psychometric standards. Items within a scale should measure coherent underlying constructs (internal consistency[26,27]), the items should exhibit interpretable factor structure (factorial validity[28,29]), and scores should predict meaningful real-world outcomes (predictive validity[30,31]). These properties are not guaranteed by item content alone, as plausible-sounding questions may fail to measure the intended construct, and theoretically motivated groupings do not always correspond to empirical latent dimensions. Empirical validation is therefore essential before a new instrument can be used with confidence.[32,33]

Several prior studies have explored this idea by using LLMs such as GPT-2 and GPT-3 to generate personality questionnaire items. For instance, Hommel and coworkers[34] fine-tuned GPT-2 on a large personality item bank to generate construct-specific items for Big Five traits. Their approach demonstrated syntactic and semantic plausibility but focused primarily on surface quality rather than empirical validity. Götz and colleagues[35] developed the Psychometric Item Generator, which generates items from construct labels alone without grounding generation in specific theoretical source texts. Hernandez and Nie[36] trained language models to predict inter-item correlations, showing that AI models can capture structural relationships between items, although their approach relied on extensive training on existing datasets rather than generating items from novel sources. Lee and colleagues[37] used prompt-based GPT-3 to generate Big Five items and compared their psychometric properties to human-authored items, but did not assess predictive validity for real-world outcomes. Collectively, these studies demonstrate the feasibility of LLM-assisted item generation. However, they generally do not ground item generation in explicit theoretical source texts beyond construct labels, make limited use of LLMs' capacity to evaluate item quality, and provide limited evidence regarding the predictive utility of the resulting questionnaires.

Here, we introduce an LLM-based pipeline for generating, validating, and evaluating personality questionnaires from textual sources, including predicting response patterns prior to data collection and assessing predictive utility for life outcomes. We hypothesized that GPT-4 could anticipate responses because it captures statistical regularities in language that align with human response patterns. This approach directly tests whether LLMs capture statistical regularities that reflect meaningful psychological structure, and whether such structure can be harnessed to build valid and useful personality assessments.

The present study pursues two related goals. First, we develop and validate a methodological pipeline for transforming pre-existing textual corpora into psychometrically evaluated personality questionnaires using LLMs. Second, we use this pipeline to examine whether questionnaires derived from conceptually distinct sources differ in their psychometric properties and predictive validity. To this end, we generated personality questionnaires from two sources: the DSM-5 personality disorder descriptions,[38] and an astrological typology,[39] administering them alongside the established Big Five Inventory (BFI[40]). These sources vary systematically in their theoretical grounding, allowing us to compare structured, scientifically grounded trait descriptions with structured but non-scientific ones.

The choice of DSM-5 descriptions is theoretically grounded in the DSM-5 alternative model for personality disorders (AMPDs), which defines personality disorders as characterized by "impairments in personality functioning and pathological personality traits"[38] (page 761) and explicitly describes these traits as "maladaptive variants of the five domains of the extensively validated and replicated personality model known as the Big Five, or Five Factor Model of personality" (page 773). DSM-5 personality disorder descriptions therefore provide a valid and theoretically grounded source of trait-relevant language, capturing maladaptive or extreme variants of personality traits without reducing them to diagnostic categories. It should be noted, however, that the DSM-5 categorical disorder descriptions and the AMPD trait dimensions, while related, are not equivalent, as each disorder description blends multiple trait-relevant features rather than mapping onto a single dimension. In doing so, we invert the logic of the AMPD, which mapped Big Five traits onto personality disorders, by instead mapping disorder descriptions back onto underlying personality dimensions.

Astrology is included as a contrast case designed to probe the role of source structure independent of scientific grounding. Astrological personality descriptions have been culturally elaborated over centuries and often resemble everyday trait-descriptive language at the surface level, yet their organization into zodiac signs and elemental categories (fire, earth, air, water) carries no empirical or scientific basis. This allows us to distinguish item-level descriptive content from the validity of the underlying organizational framework, and to test whether the psychometric coherence and predictive utility of LLM-generated questionnaires depend on theoretically grounded constructs or can emerge from structured language alone.

RESULTS

Automatic generation of personality questionnaires using LLMs

To generate items suitable for personality assessment, we used OpenAI's GPT-4 model[41] with a structured, few-shot learning prompt approach (Figure 1). We provided example items from the validated (BFI-44[40]), a questionnaire consisting of 44 items measuring the five personality dimensions. This guided the model to produce items consistent with standard personality questionnaire formats, specifically Likert-scale statements rated from 1 to 5. The DSM-5[38] describes ten personality disorders as maladaptive extremes of normal traits, grouped into three higher-order clusters (A, B, and C) reflecting broader personality patterns. We aimed to generate items representing mild, non-pathological expressions of these traits. For the astrology-based questionnaire, items were generated to reflect commonly attributed personality characteristics associated with 12 zodiac signs. These are grouped into four higher-order clusters corresponding to the classical elements (Water, Fire, Air, Earth), capturing shared traits of the zodiac signs within each element.[39] Using our structured GPT-4 few-shot procedure, we generated 50 items from DSM-5 personality disorder descriptions (5 items per disorder: 15 items for Cluster A, 20 for Cluster B, 15 for Cluster C) and 60 items from astrological zodiac sign descriptions (5 items per zodiac sign, 15 items per classical element).

Please cite this article in press as: Monsa et al., Generating and analyzing personality questionnaires using large language models, iScience (2026), https://doi.org/10.1016/j.isci.2026.116909

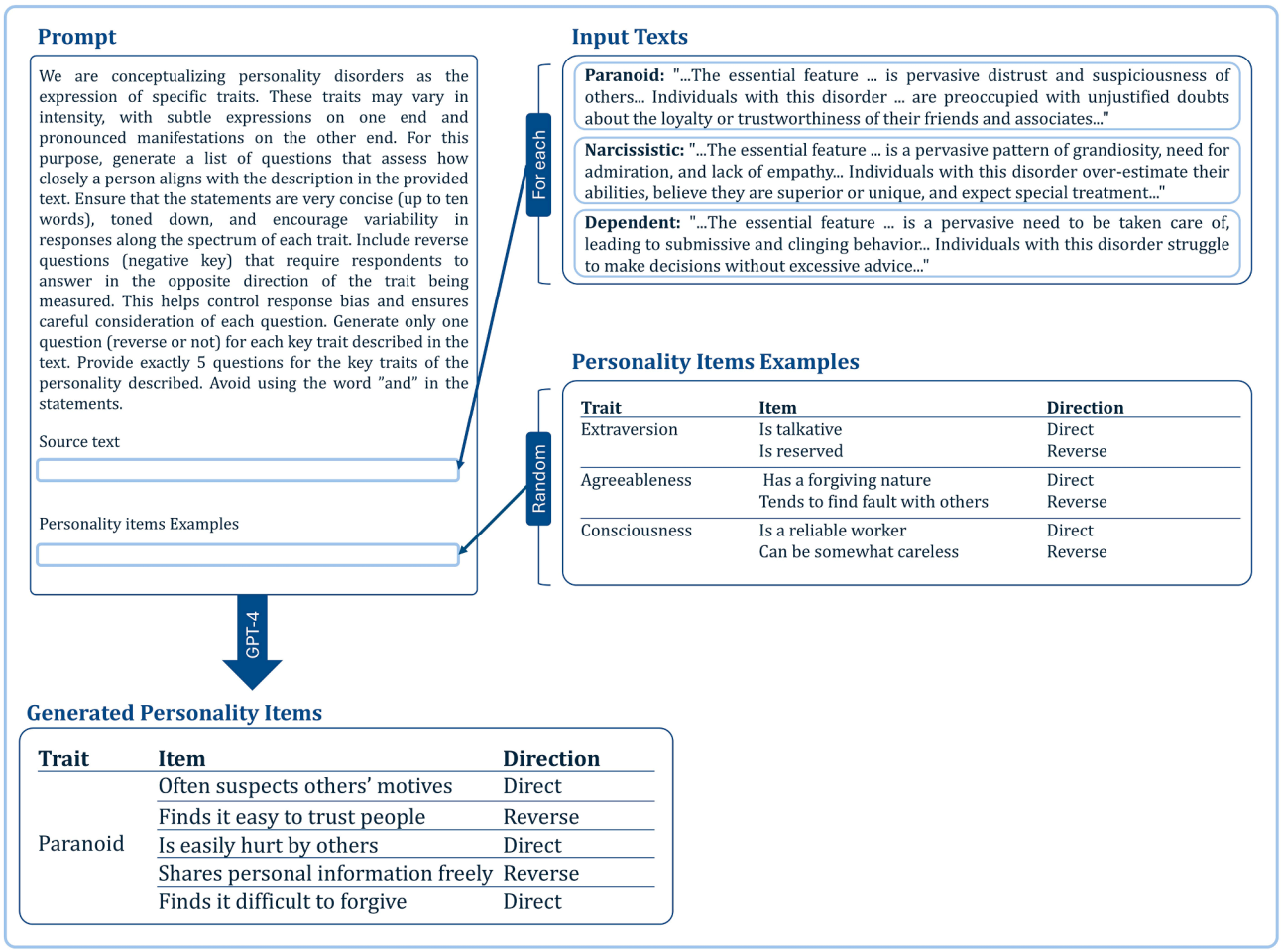


Figure 1. A schematic representation of the process for generating personality questionnaire items using GPT-4 from texts

Schematic illustration of the prompting workflow used to generate personality assessment items based on textual descriptions, using GPT-4. In this example, representative excerpts from DSM-5 personality disorder descriptions (e.g., Paranoid, Narcissistic, Dependent) are used as input (upper row, right), along with sample items from standard personality inventories (middle row, right) to guide tone, structure, and item direction. The prompt (upper row, left) instructs GPT-4 to generate five concise items per trait, capturing distinct aspects of each description and balanced across direct and reverse-keyed phrasing to reduce response bias. Output items for the Paranoid trait are shown as an example (lower row). The same procedure was applied to other textual sources, including astrology descriptions, a user manual, and a passage from The Lord of the Rings.

All generated items were evaluated using GPT-4 for alignment with the source descriptions and overall quality (see STAR Methods). For the DSM-based questionnaire, 38 out of 50 items had perfect alignment with their intended personality disorder, and after revisions, 40 items achieved perfect alignment, resulting in a total normalized content validity score of 0.96. At the end of the generation and refinement process, item quality evaluation indicated that the generated items were clear, coherent, and suitable for Likert-scale responses. Structural and linguistic evaluation indicated that the generated items were of high quality and comparable to standard personality items across multiple indices (Table 1). DSM-based items were slightly longer and less readable than BFI items. Grammatical acceptability was high across all three sets, with LLM-generated items matching or exceeding the BFI benchmark. Items reflected both positively and negatively keyed phrasing, with the DSM-based questionnaire having the most balanced keying (54% positive, 46% negative), consistent with the need to represent both pathological and healthy poles of each trait (for full item-level metrics see Table S1). Overall, GPT-4-generated items demonstrated high specificity and coherence with their source descriptions, producing two distinct, psychometrically promising item sets. Full item lists and response statistics are provided in Tables S2 and S3.

Effects of non-personality texts on generated items

As an additional control, we applied the same generation procedure to two texts unrelated to personality: an electric oven user manual[42] and a landscape description from *The Lord of the Rings*.[43] Despite the lack of personality-related content, the model consistently generated items with a recognizable personality-like structure, often using trait-like adjectives and framing statements in the familiar style of personality questionnaires (Figure 2). However, while the items were structurally consistent,

Please cite this article in press as: Monsa et al., Generating and analyzing personality questionnaires using large language models, iScience (2026), https://doi.org/10.1016/j.isci.2026.116909

Table 1. Structural and linguistic characteristics of items across the three questionnaires

Structural metric	BFI	DSM-based	Astrology-based
Average words per item (SD)	5.01 (2.80)	6.45 (2.82)	5.65 (2.82)
Readability ease (SD)	72.27 (14.87)	62.30 (14.17)	67.62 (15.02)
Item keying			
Positive-keyed items	28 (64%)	27 (54%)	40 (67%)
Negative-keyed items	16 (36%)	23 (46%)	20 (33%)
Linguistic acceptability			
Grammatically acceptable items	88.64%	100.00%	98.33%

Readability ease assessed using the Flesch Reading Ease score. Grammatical acceptability assessed using RoBERTa-base fine-tuned on the corpus of linguistic acceptability (CoLA) (for details see Supplementary Methods and Table S4; SD = standard deviation.

they were often semantically incoherent or irrelevant. Therefore, we determined they did not meet the minimum threshold for inclusion in the human response phase. These results suggest that the LLM integrated content elements and contextual cues from the input text while adhering to the structural format of personality items. These results reflect both the model's generative flexibility and the biases it introduces into the generation process.

LLM-based correlation and average response estimation

To evaluate whether GPT-4 could anticipate patterns in questionnaire responses *a priori*, we asked the model to estimate both the average response (mean score) for each item and the Pearson's correlations between item pairs, prior to collecting participant data (full prompts shown in Figures S2 and S3). After administering the questionnaires to 600 participants (recruited online; see STAR Methods), we compared GPT-4's estimations to the empirical results. The model's estimated mean responses were strongly correlated with actual participant responses: r = 0.71, 95% CI [0.56, 0.82] for DSM-based items, and r = 0.85, 95% CI [0.77, 0.91] for astrology-based items (both $p < 0.0001$; Figure 3). In line with these results, ICC analyses indicated substantial agreement (ICC(2,1) = 0.69 for the DSM-based questionnaire; ICC(2,1) = 0.84 for the astrology-based questionnaire). The mean absolute error (MAE) for average response prediction was 0.40 for DSM-based items and 0.27 for astrology-based items, corresponding to 8% and 5.4% of the total scale range, respectively. Estimated correlations also closely matched observed item correlations: for the DSM-based questionnaire, r = 0.74, 95% CI [0.69, 0.78]; for the astrology-based questionnaire, r = 0.69, 95% CI [0.64, 0.73] (both $p < 0.001$; Figure 4). The corresponding MAE for correlation prediction was 0.17 for both questionnaires, corresponding to approximately 8.5% of the total possible correlation range. These results suggest that GPT-4 can infer both response distributions and correlational structure directly from item text, reflecting the extent to which LLMs capture population-level response patterns, and indicating their potential utility as a pre-screening tool during early-stage questionnaire development.

Internal consistency

To assess whether items within each trait measured a coherent underlying construct, we estimated internal consistency using Cronbach's alpha at both the facet and cluster levels. For the LLM-generated questionnaires, "facets" refer to the individual personality disorders in the DSM-based questionnaire (10 disorders) and the individual zodiac signs in the astrology-based questionnaire (12 signs), whereas clusters correspond to broader groupings (e.g., DSM Clusters A–C, FFM five factors, astrology four classical elements). For the BFI, facet-level alphas ranged from 0.41 to 0.85, with high reliability across the five broader factors (α = 0.80–0.88). The DSM-based model showed facet-level alphas of 0.40–0.65, with one exception: the obsessive-compulsive personality disorder facet, which showed $\alpha = 0$. At the broader level, cluster-level consistency was higher: Clusters A and B both showed $\alpha = 0.76$, and Cluster C $\alpha = 0.66$. This pattern suggests that GPT-generated items based on diagnostic descriptions captured meaningful shared variance, particularly at the cluster level. In contrast, astrology-based facets showed weaker internal consistency ($\alpha = -0.46$–0.62), and cluster-level alphas for the four classical elements were lower (α = 0.45–0.63). This likely reflects limited conceptual coherence in the underlying astrological constructs rather than a failure of the item-generation process (Tables S4 and S5). We note that Cronbach's alpha is sensitive to scale length, so differences in internal consistency across facets may partly reflect differences in item count rather than construct coherence.

Confirmatory factor analysis (CFA)

To evaluate the correspondence between theoretical and observed structure, confirmatory factor analyses (CFA) were conducted for each questionnaire separately. Each model specified the hypothesized grouping structure (BFI: five factors; DSM-based: three clusters; astrology-based: four classical elements), with items loading only on their designated factor and latent factors allowed to correlate. For the BFI, model fit was moderate given the complexity of omnibus personality models (CFI = 0.72, TLI = 0.70, RMSEA = 0.078), consistent with known limitations of CFA under strict simple-structure assumptions in personality assessment.[44] The DSM-based and astrology-based models showed weaker correspondence with the observed covariance structure (DSM: CFI = 0.53, TLI = 0.50, RMSEA = 0.083; astrology: CFI = 0.39, TLI = 0.36, RMSEA = 0.089). These results indicate that the theoretical groupings only partially reproduce the observed item covariance structure under CFA constraints.

Exploratory factor analysis (EFA)

To complement the confirmatory analyses, exploratory factor analysis (EFA) was conducted without imposing theoretical constraints. The number of factors was determined using permutation-based parallel analysis, and factor extraction was performed using principal axis factoring with oblimin rotation.

For the BFI, a six-factor solution explained 44.9% of the variance and closely reproduced the canonical Big-Five structure.

Please cite this article in press as: Monsa et al., Generating and analyzing personality questionnaires using large language models, iScience (2026), https://doi.org/10.1016/j.isci.2026.116909

Example generation of personality questionnaire items from non-personality related texts

Oven Manual Source Text

"Your appliance comes with a range of accessories. Below is an overview of the included accessories and how to use them correctly...

- **Wire Rack:** Suitable for ovenware, cake tins, ovenproof dishes, roasts, grilled items, and frozen meals.
- **Universal Pan:** Ideal for moist cakes, pastries, frozen meals, and large roasts...
- **Grill Tray:** Used for grilling in place of the wire rack or as a splatter guard...
- **Meat Thermometer:** Allows precision roasting to your exact requirements..."

Mordor Source Text

"All was dark about it, earth and sky alike. Sickly vapours hung like veils in the heavy air... the lasting mark of the dark lord's reign.
Before them, dark in the dawn, the great mountains reached up to roofs of smoke and cloud... broken hills now scarce a dozen miles away.
The gasping pools were choked with ash and crawling muds, sickly white and grey, as if the mountains had vomited the filth of their entrails...
Here nothing lived, not even the leprous growths that feed on rottenness... a land defiled, diseased beyond all healing.
All was ominously quiet. The light was no more than that of dusk at a dark day's end... now lit again beneath with a sullen glow of red.
The bitter mountains of iron rose like fangs... only dust and smoke and ash beneath the sullen clouds."

Generated Personality Items from Oven Manual

Adjective	Statement	Key
Meticulous	Is meticulous with cookware placement	Direct
Flexible	Struggles with accessory variability	Reversed
Observant	Notices minor design changes quickly	Direct
Organized	Has well-arranged kitchen items	Direct
Innovative	Rarely tries new cooking setups	Reversed

Generated Personality Items from Mordor Description

Adjective	Statement	Key
Desolate	Feels a sense of emptiness	Direct
Grim	Rarely sees things positively	Direct
Oppressive	Feels burdened by surroundings	Direct
Sickly	Often feels unhealthy	Direct
Barren	Lacks vitality or growth	Reversed

Figure 2. Example generation of personality questionnaire items from non-personality-related texts

Shown are example excerpts (upper row) from an electric oven user manual (left) and a literary description of Mordor's landscape from *The Lord of the Rings* (right), used as input sources for item generation (text excerpts are partial and selected for illustrative purposes). GPT-4-generated personality items based on each source are shown in the lower row. Note the anchoring to the source in both textual corpuses.

Of its five dimensions, Neuroticism, Openness, Conscientiousness, and Extraversion emerged as distinct factors, while Agreeableness split into two components reflecting agreeableness versus low antagonism. A five-factor solution also recovered the standard Big-Five structure, with the sixth factor reflecting a finer-grained separation within Agreeableness.

For the DSM-based questionnaire, the seven-factor solution (35% variance explained) yielded interpretable clinical dimensions including social detachment, attention-seeking, emotional instability, impulsivity/low self-esteem, low empathy, disinhibition, and paranoid hostility.

For the astrology-based questionnaire, the seven-factor solution (36% variance explained) did not preserve the theoretical element structure. Instead, items organized around broader personality dimensions, with the first four factors mapping onto core Big-Five domains (Agreeableness, Conscientiousness, Extraversion, and Openness). Remaining factors reflected more diffuse affective and interpersonal tendencies.

Overall, the EFA results highlight distinct patterns across the three instruments. The BFI largely reproduces its expected hierarchical structure. The DSM-based questionnaire yields a more granular but clinically interpretable factor structure than its theoretical three-cluster grouping. The astrology-based questionnaire shows no meaningful alignment with its imposed elemental structure and instead maps onto broad Big-Five dimensions, consistent with the theoretically imposed rather than empirically derived nature of the elemental groupings (for factor loadings see Tables S6–S8).

Comparison of model performance in predicting life outcomes

A central goal of personality models is to predict meaningful life outcomes. We compared the predictive performance of the BFI-44, DSM-based, and astrology-based personality questionnaires across a range of life outcomes using 10-fold cross-validation in a sample of participants who completed all three personality assessments. All models were trained using individual questionnaire items as predictors rather than aggregated factor scores, allowing us to assess the predictive utility of the full item set for each personality model. For ordinal outcomes, performance was evaluated using quadratic weighted kappa (QWK). Performance was comparable across the three models for the large majority of outcomes. The only significant difference was observed for education level, where the astrology-based questionnaire outperformed the BFI-44 questionnaire (p value <0.05, FDR-corrected, Wilcoxon signed-rank test following a significant Friedman test). No significant difference was found between the astrology-based and DSM-based questionnaires for this outcome. For the non-ordinal categorical outcomes, including employment status, marital status, and political orientation, performance was evaluated using Cohen's kappa. The Friedman test revealed no significant differences between questionnaires for any of these outcomes (all p values >0.05). Overall, the three questionnaires performed similarly across almost all life outcomes, with the only exception being a modest advantage of the astrology-based model over the BFI-44 in predicting education level (Figure 5; Figure S4; Table S9).

Please cite this article in press as: Monsa et al., Generating and analyzing personality questionnaires using large language models, iScience (2026), https://doi.org/10.1016/j.isci.2026.116909

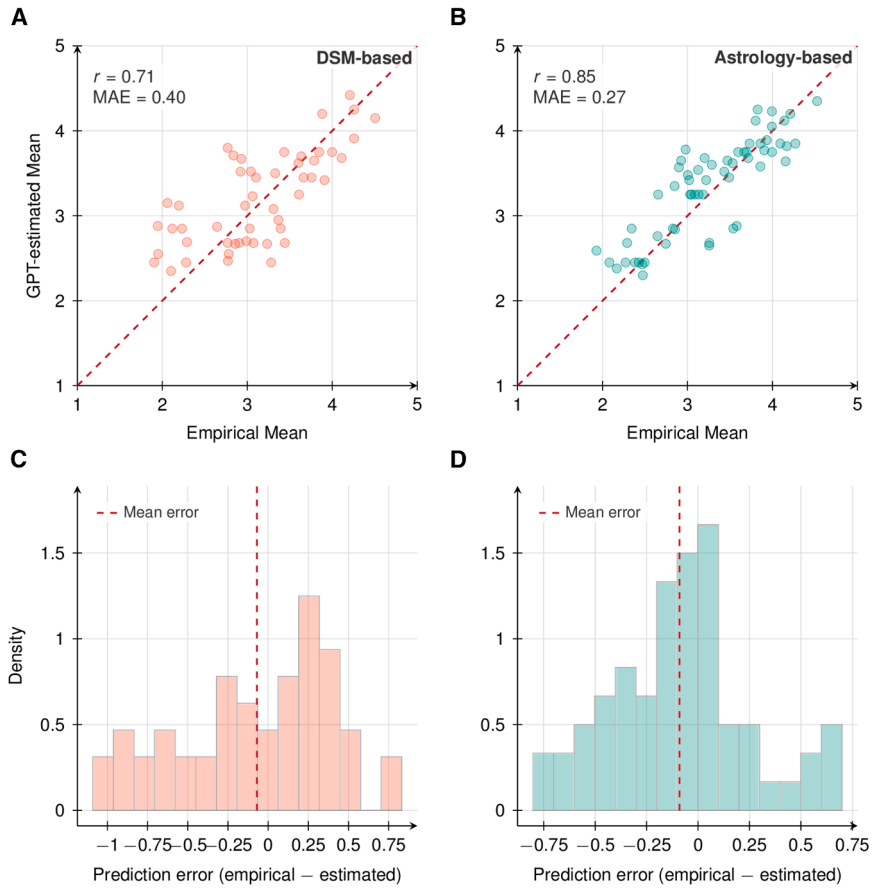


Figure 3. GPT-4 estimation performance for average item responses in DSM- and astrology-based questionnaires

(A and B) Scatterplots compare GPT-4 estimated mean responses with empirical mean responses for DSM-based (A) and astrology-based (B) items. Each point represents a unique item. The red dashed line represents perfect estimation (y = x). Points closer to the dashed line indicate more accurate mean response predictions by GPT-4. Points above the line indicate underestimation of the empirical mean response, while points below the line indicate overestimation. The distance from the line corresponds to the magnitude of the prediction error.

(C and D) Density histograms of prediction errors (empirical mean − GPT-4 estimated mean) for DSM-based (C) and astrology-based (D) items. Red dashed lines mark the mean errors (−0.07 and −0.09, respectively). Errors centered around zero indicate unbiased predictions; positive errors reflect underestimation, and negative errors reflect overestimation.

At the scale level, significant differences emerged for mental health outcomes (Friedman tests, all p values <0.01), with post-hoc comparisons revealing that both the DSM-based questionnaire and the BFI significantly outperformed the astrology-based questionnaire (all FDR-corrected p values <0.05), while not differing significantly from each other. This disadvantage for the astrology-based questionnaire was specific to mental health outcomes. For well-being, driving behavior, substance use, and demographic outcomes, all three questionnaires performed comparably at the scale level. No significant differences were observed between models for most outcome domains, consistent with the item-level results reported above (for detailed statistics see Table S10).

We additionally compared item-level versus factor-score model performance to examine whether using individual items rather than aggregated factor scores improves predictive accuracy. Item-level models significantly outperformed factor-score models for all three questionnaires (Wilcoxon signed-rank tests across all outcomes per model, all $p < 0.001$), with item-level models performing better in the large majority of comparisons. Performance was consistently higher for item-level models than for factor-score models for the BFI (0.24 vs. 0.14; Δ = 0.10), the DSM-based questionnaire (0.24 vs. 0.16; Δ = 0.08), and the astrology-based questionnaire (0.23 vs. 0.12; Δ = 0.11). The largest advantage was observed for the astrology-based questionnaire, consistent with the relatively low internal consistency of its factor structure, suggesting that aggregation into factor scores may reduce predictive information when the underlying constructs are less coherent (Table S11).

This analysis can be interpreted as a direct comparison of how much real-world predictive information is encoded in each questionnaire at the item level. The overall similarity in performance suggests that all three instruments, despite their different theoretical origins, capture comparable amounts of variance associated with external life outcomes.

DISCUSSION

Our results suggest that LLMs (here: GPT-4) show promise for generating personality questionnaire items based on textual sources (here: DSM-5 and an astrology textbook). Trained on large-scale textual data that includes descriptions of human behavior and personality, LLMs appear capable of approximating aspects of expert reasoning in generating questionnaire items that are clear, conceptually consistent with their source material, and aligned with intended traits. The automated item generation (AIG) process yielded items that were judged as theoretically interpretable, as verified by GPT-4-based content validation and further confirmed through manual inspection. Moreover, GPT-4 was able to estimate inter-item correlations and mean response prior to data collection, based solely on item content, suggesting that LLMs encode information about population-level response tendencies. In addition to generation, LLMs can be used to examine psychometric properties of the resulting questionnaire, supporting further refinement and precision. Finally, predictive performance across various life outcomes was broadly comparable between questionnaires

Please cite this article in press as: Monsa et al., Generating and analyzing personality questionnaires using large language models, iScience (2026), https://doi.org/10.1016/j.isci.2026.116909

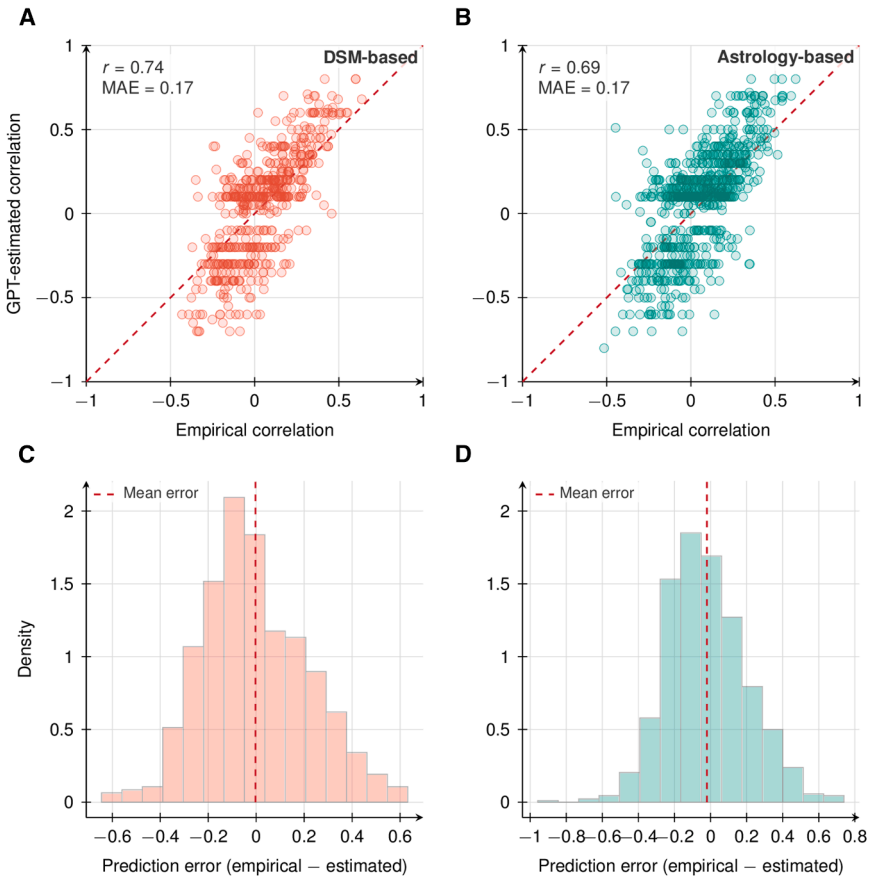


Figure 4. GPT-4 correlation estimation performance for the DSM- and Astrology-based questionnaires

(A and B) Scatterplots compare empirical inter-item correlations with GPT-4 estimated correlations for DSM-based (A) and astrology-based (B) questionnaires. Each point represents a unique item pair. The red dashed line represents perfect estimation (y = x). Points closer to the dashed line indicate more accurate correlation predictions by GPT-4. Points above the line indicate underestimation of the empirical correlation, while points below the line indicate overestimation. The distance from the line corresponds to the magnitude of the prediction error.

(C and D) Density histograms of prediction errors (empirical correlation − GPT-4 estimated correlation) for DSM-based (C) and astrology-based (D) questionnaires. Red dashed lines mark the mean errors (−0.003 and −0.02, respectively). Errors centered around zero indicate unbiased predictions; positive errors reflect underestimation, and negative errors reflect overestimation.

derived from a major psychiatric corpus (DSM-5), a popular astrological text, and the widely accepted FFM, highlighting the flexibility of LLM-generated item sets across theoretical frameworks. These results are further discussed later in discussion.

The successful generation of personality questionnaire items from both the DSM-5 and astrological texts extends prior work on AIG using LLMs. Previous work on the LLM-based generation of personality questionnaire items has reported mixed psychometric properties, particularly with earlier GPT-2/3-based approaches and fine-tuned models.[34–37] In contrast to these approaches that primarily depend on fine-tuning LLMs with psychometric datasets, our study uses a prompt-based strategy that integrates personality descriptions ranging from clinical frameworks to less conventional sources such as astrology, without additional model fine-tuning. This approach demonstrates that high-quality items can be generated across different theoretical frameworks when appropriate contextual information is provided to the model.

More broadly, this work contributes to the growing integration of LLMs in psychological research, where these models are increasingly being used across multiple domains including cognitive modeling, clinical assessment, mental health screening, and digital interventions.[45,46] This is of particular relevance in light of recent developments in the field of neurodegenerative disorders, such as Alzheimer's disease, in which AI-based tools have enabled clinicopathological predictions based on subtle cognitive and behavioral changes.[47–50] Similarly, the approach proposed here may support the development of such tools for mental health conditions, including affective, post-traumatic, and personality disorders.

Likewise, LLMs and embedding models have been successfully used to detect depression and suicide risk from patient narratives,[51] to classify symptoms of comorbid depression and anxiety in clinical message data,[52] and to identify depression from user-generated diary text.[53] Our results highlight the potential of LLMs to complement traditional empirical methods by rapidly generating, refining, and validating theory-driven instruments across diverse psychological frameworks.

Interestingly, GPT-4 produced personality-like items even when prompted with non-personality-related source texts, such as an electric oven user manual or a literary landscape description. Despite being given no explicit personality trait descriptions, the model consistently produced items in a personality questionnaire format. This may reflect the model's internal knowledge of what constitutes personality. For example, items such as "Rarely sees things positively" or "Has well-arranged kitchen items" are semantically grounded in familiar personality traits such as pessimism or organization, even if derived from unrelated inputs such as literary landscapes or appliance manuals. At the same time, the source text clearly influences the specific framing of items, sometimes resulting in overly narrow or specific expressions of traits.

These findings suggest that input texts vary systematically in their suitability for LLM-based personality questionnaire generation. Suitable source materials appear to be those that provide semantically grounded descriptions of human behavior, traits, or psychological dispositions, which can be abstracted into

Please cite this article in press as: Monsa et al., Generating and analyzing personality questionnaires using large language models, iScience (2026), https://doi.org/10.1016/j.isci.2026.116909

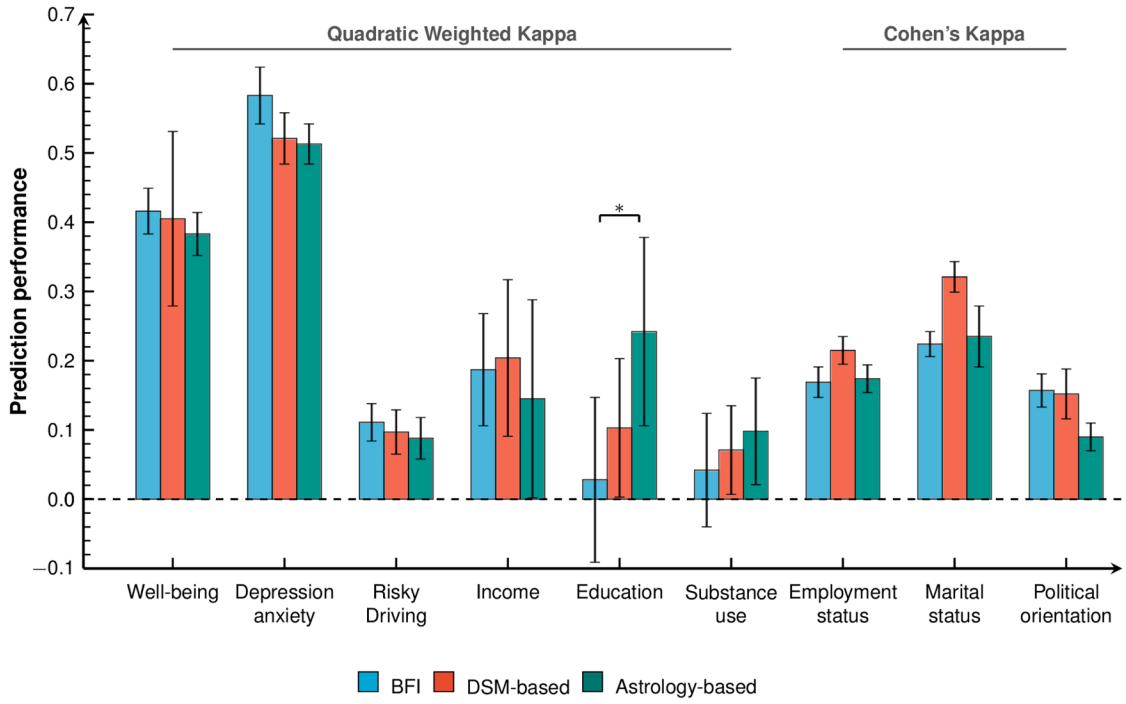


Figure 5. Predictive performance of life outcomes for the three personality models

Bars represent the prediction performance of the BFI (blue), DSM-based (orange), and astrology-based (green) models. Performance is measured using quadratic weighted kappa (QWK) for ordinal outcomes (left) and Cohen's kappa for categorical outcomes (right), as indicated by the labels above the plot. For multi-item outcomes, such as well-being (4 items), depression/anxiety (4 items), risky driving (9 items), and substance use (2 items), bars represent the average QWK across items. Outcomes include psychological measures (e.g., well-being, anxiety, risky driving), health behaviors (substance use), and demographic variables (education, income, employment status, marital status, and political orientation). Error bars indicate standard deviations across cross-validation folds. An asterisk marks a significant difference between models for education level ($p < 0.05$, FDR-corrected Wilcoxon signed-rank test following a significant Friedman test).

stable constructs. Importantly, suitable texts should also provide broad coverage of personality traits and behavioral tendencies, such that multiple distinct traits or behavioral tendencies can be represented rather than a narrow subset of behaviors. In contrast, texts lacking behavioral grounding or comprehensive coverage may still elicit well-formed questionnaire-like structures due to the model's strong inductive bias toward psychometric formats, but are less likely to yield psychologically meaningful or construct-valid items. This pattern highlights a distinction between structural priors inherent to the language model and semantic constraints provided by the input text in determining item validity.

Another key aspect of questionnaire quality is internal consistency, which reflects the degree to which items within a factor or cluster measure a coherent construct. As expected, internal consistency at the facet level was lower for both the BFI (individual facets) and the DSM-based questionnaire (personality disorders), while consistency was higher when items were aggregated into broader clusters or factors. At the factor level, the BFI showed high internal consistency, consistent with its extensive refinement. The DSM-based questionnaire exhibited good and acceptable consistency across the three higher-order clusters (A, B, C), especially for exploratory research purposes.[54,55] The astrology-based questionnaire, in contrast, lacked clear internal consistency across all four classical elements. Importantly, this does not reflect a failure of item generation, since all items were produced using the same LLM procedure and anchored in their source text. Rather, the low internal consistency mirrors the conceptual structure of astrology itself, which does not group traits according to coherent latent dimensions.

Across the three instruments, the CFA and EFA results highlight a distinction between theoretically imposed structure and empirically emergent organization. The BFI showed the closest correspondence to its expected structure, consistent with prior psychometric work. In contrast, the DSM-based and astrology-based questionnaires showed weaker fit under confirmatory models, with fit indices falling below acceptable thresholds. Notably, some degree of misfit is expected when complex trait structures are constrained by zero cross-loadings, as is inherent in standard CFA.[44,56] The EFAs provide additional insight into the underlying structure. For the DSM-based questionnaire, exploratory factors were partially interpretable in terms of maladaptive trait content, showing partial convergence with broad dimensions conceptually related to the AMPD (e.g., affective instability, detachment, antagonism, and disinhibition), but did not reproduce the intended three-cluster model. This pattern is consistent with the distinction between DSM-5 categorical

Please cite this article in press as: Monsa et al., Generating and analyzing personality questionnaires using large language models, iScience (2026), https://doi.org/10.1016/j.isci.2026.116909

disorder descriptions and the AMPD trait dimensions, in which each categorical description blends multiple trait-relevant features rather than mapping onto a single dimension, which may limit recovery of a clean three-factor structure. For the astrology-based questionnaire, the imposed elemental structure was not supported, and items instead reorganized around broader trait dimensions broadly consistent with the Big Five. These findings suggest that LLM-generated item structure depends strongly on the representational properties of the source text, and that meaningful item content does not necessarily imply recovery of the intended latent model. Future work employing exploratory structural equation modeling (ESEM) may better accommodate the cross-loading structure inherent in these item sets.

One surprising finding was the LLM's ability to estimate the average response that human participants would give to individual personality items without any examples. This is particularly interesting given that personality questionnaires aim to differentiate individuals based on their unique responses. An inspection of the response data revealed that item responses showed moderate variability across participants, suggesting that GPT-4's success in predicting average responses reflects its understanding of population-level response distributions rather than a lack of variability in the items. Importantly, this result should not be interpreted as evidence of questionnaire validity, but rather as an indication that LLMs capture general response patterns in human populations. One practical implication of this approach is its potential use as a pre-screening tool during questionnaire development. However, the accuracy of such predictions is likely to depend on the match between the population implicitly assumed by the model and the sampled population. An alternative approach in the literature is to simulate synthetic individuals or personas and aggregate their responses to estimate population patterns.[57–59] In contrast, the present study directly queries the model for aggregate statistics such as expected means and inter-item correlations. Future work should examine the conditions under which direct estimation and persona-based simulation yield equivalent or divergent results.

In addition, GPT-4 predicted inter-item correlations with notable accuracy even prior to data collection. Previous research has achieved high accuracy in this task using a paired DistilBERT architecture fine-tuned on over 3.4 million correlations from large personality datasets.[36] Notably, our study demonstrates that even a general-purpose LLM, without any fine-tuning for this task, can still achieve substantial accuracy in predicting item correlations, relying only on its broad language understanding and a small number of example items. GPT-4's ability to predict correlations likely emerges from its implicit understanding of language, human psychology, and population-level response patterns.

These results align with recent findings showing that LLMs can accurately predict neural representations in the human language system,[25,60] and that they segment narrative events in ways consistent with human perception.[61] Together, these results suggest that LLMs' internal representations capture meaningful aspects of how humans process and respond to language, including high-level conceptual and relational structures. These findings have important implications for psychometric practice, suggesting that even general-purpose LLMs can provide preliminary psychometric estimates before human data collection, enabling iterative optimization of instruments and accelerating scale development.

While GPT-4 demonstrates notable accuracy in capturing the internal structure of personality questionnaires, a key question remains: how much do these items capture meaningful individual differences? To address this, we used predictive validity for life outcomes as an illustrative metric. This was chosen not because the BFI or our LLM-generated questionnaires were designed specifically for prediction, but simply as a quantitative way to assess whether the items reflect variance that relates to real-world outcomes. Evaluating predictive validity in this way provides a concrete benchmark for comparing instruments and offers an indirect measure of questionnaire quality, since higher predictive performance suggests that the items capture meaningful and relevant psychological variance. Since personality models are used to predict people's choices, feelings, and behaviors,[62] examining their predictive power helps evaluate the practical utility of the questionnaires.

A key theoretical implication of our findings is that internal structural validity and predictive validity represent partially separable aspects of construct validity. Across analyses, we observed a dissociation between latent structure (as assessed via CFA and EFA) and external predictive performance. In particular, the DSM-based questionnaire showed limited evidence of a coherent latent structure under CFA and a more fragmented factor structure under EFA, yet still exhibited substantial predictive validity at the aggregated level. This suggests that predictive signal can be distributed across heterogeneous item content even in the absence of strong factorial coherence.

A broader theoretical implication of these findings is that different conceptual frameworks (DSM-based, astrology-based, and Big Five) show comparable explanatory power for life outcomes. This raises the possibility that personality prediction may be more sensitive to item-level linguistic content than to the theoretical framework from which items are derived. Consistent with this interpretation, item-level models outperformed aggregated factor-score models across all three questionnaires.[63] Across all questionnaires, diverse, high-signal items yielded predictive accuracy superior to full factor structures. Even astrology-derived items, despite lacking conceptual validity, showed meaningful predictive power. These results suggest that factor-level aggregation in the BFI and DSM-based questionnaires contributes little to predictive utility, and that predictive performance is primarily driven by a set of distinct, high-signal items. Importantly, this demonstrates that the LLM-generated items capture meaningful, high-signal variance.

A complementary pattern emerged for the astrology-based questionnaire, which further highlights the role of aggregation. Astrological personality descriptions have been refined over centuries of cultural use and overlap substantially with everyday trait-descriptive language, which explains why individual items retained measurable predictive signal. However, aggregation according to theoretically imposed elemental groupings reduced predictive performance. This is consistent with a mismatch between the imposed structure and the empirical covariance structure, as the assignment of items to fire, earth, air, and water

categories evidently does not reflect the actual latent dimensions recovered empirically. In contrast, the BFI and DSM-based instruments benefited more from aggregation at the scale level, particularly for psychologically complex outcomes, suggesting that structural coherence may aid in preserving predictive information when constructs are well-aligned with empirical dimensions.

Importantly, item-level analyses revealed that predictive information is not solely dependent on latent structure, but can also arise from the semantic content of individual items, which remains informative even when higher-order structure is weak or misaligned. Taken together, these findings align with the view that structural model fit should not be used as the sole criterion for construct validation.[56] Instead, internal structure (CFA/EFA results) and external prediction reflect partially separable components of construct validity, such that improvements in one domain do not necessarily translate into improvements in the other.

In conclusion, our findings point to a shift in how personality evaluation and psychological questionnaires in general can be approached. Our approach enables rapid, transparent, and replicable generation and evaluation of personality items without requiring extensive data collection, expert input, or model fine-tuning. This allows researchers to quickly operationalize diverse theoretical constructs and iteratively refine instruments before investing in costly human data collection. LLMs' ability to estimate psychometric properties and predict item responses in advance can further accelerate scale development and reduce the time and resources required in traditional psychometric work. Importantly, while additional psychometric refinement remains necessary, the results suggest that efficient LLM-based item generation can produce informative item sets with meaningful predictive signal: the broadly comparable predictive performance of astrology-based, DSM-based, and BFI questionnaires suggests that LLM-generated items capture meaningful psychological variance even when derived from diverse or unconventional sources. These findings open new possibilities for rapid, scalable theory testing and suggest several promising directions for future research. In particular, future work may move beyond constrained questionnaires and leverage LLM capabilities for direct personality assessment from natural language, which may capture personality more richly than traditional self-report measures.

Limitations of the study

Our study is not free of limitations. While we demonstrated the feasibility of using LLMs to generate and validate personality items across different frameworks, further work is needed to refine these models and ensure broader generalizability. First, like all LLMs, GPT-4 may reflect cultural or theoretical biases embedded in its training data, potentially influencing which traits are emphasized or how they are expressed.[64,65] Nonetheless, in the context of personality questionnaire development, the model's broad knowledge base, drawn from diverse internet sources, may help mitigate some of the limitations associated with individual human biases by offering a wider and more integrative perspective. Second, our approach was based entirely on a single model, GPT-4. While this provided a consistent framework for testing feasibility, future studies should investigate whether comparable results are obtained with other LLMs. Third, certain design choices may have influenced item properties: we used a limited set of few-shot examples (six BFI items per trait), which provided clear stylistic guidance but may have constrained semantic diversity, and we generated a fixed number of items per construct (five), which may have limited coverage of multidimensional constructs. The near-zero internal consistency of the obsessive-compulsive personality disorder facet likely reflects both its inherent multidimensionality and the challenges of capturing it in a small number of items. Fourth, our empirical data were collected through an online platform, which limits control over participant characteristics and the level of engagement with the task. Responses for both personality items and life outcomes were self-reported, introducing potential sources of noise and bias such as inattentive responding, self-presentation effects, and inaccuracies in self-assessment. However, such limitations are common in large-scale psychological research, and we employed several verification strategies to help mitigate their impact, including attention checks, response time screening, and verification that excluded participants did not differ systematically from the retained sample on demographic variables. Additionally, while self-report measures are susceptible to social desirability bias, anonymous online administration has been shown to reduce such effects relative to face-to-face assessment,[66,67] and data collected via Prolific have been shown to be of higher quality and less susceptible to response bias compared to other online platforms.[68,69] Test-retest reliability was not assessed in the current study and should be examined in future work to evaluate the temporal stability of LLM-generated questionnaires. It should be noted that GPT-4 tends to reflect Western, English-speaking populations,[64,70] and that demographic factors such as age may influence mean personality response levels.[71] Generalizability to other samples therefore requires further investigation. Additionally, all predictive analyses were conducted using internal cross-validation within a single dataset, which limits the generalizability of the findings. External validation across independent samples, diverse populations, and different cultural contexts will be necessary to establish the broader applicability of these results.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Rotem Monsa (rotem.monsa@mail.huji.ac.il).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- All data and LLM prompts are openly available in the project repository: OSF: https://osf.io/3t8zh/overview?view_only=547e7d3775c545e5aea6174c4845cc7b. The repository includes all generated item sets and analysis outputs used in this study. However, the original textual materials used as inputs for item generation (including DSM diagnostic descriptions, passages from an astrology book, and excerpts from *The Lord of the Rings*) cannot be shared due to copyright restrictions. To ensure transparency, we provide detailed descriptions of all input

sources and include prompts with placeholders indicating where copyrighted text was inserted.
- All analysis code used in this study is available in the same OSF repository.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

The study was supported by the Israel Science Foundation (732/2024). We thank Prof. Ariel Knafo-Noam and Dana Katsoty for their valuable input on the theoretical design and implementation of the questionnaires, and Fatima Jaber for her assistance in formatting the questionnaires in QuestionPro.

AUTHOR CONTRIBUTIONS

Conceptualization, R.M., A.Z., and S.A.; methodology, R.M., A.Z., and S.A.; investigation, R.M.; data curation, R.M.; formal analysis, R.M.; software, R.M. and A.Z.; visualization, R.M.; writing – original draft, R.M., A.Z., and S.A.; writing – review and editing, R.M., A.Z., and S.A.; supervision, A.Z. and S.A.; funding acquisition, A.Z. and S.A.

DECLARATION OF INTERESTS

The authors declare no competing interests.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the author(s) used GPT-4o and Claude 3.5 Sonnet in order to check English expressions. After using this tool or service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Participants
- METHOD DETAILS
  - Personality item generation
  - LLM-based item validation
  - Correlation estimation
  - Average response estimation
  - Human participants study
  - Confirmatory and exploratory factor analyses (CFA and EFA)
  - Predictive modeling of life outcomes
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2026.116909.

Received: January 26, 2026
Revised: April 28, 2026
Accepted: July 7, 2026

REFERENCES

1. Goldberg, L.R. (2013). An alternative "description of personality": The Big-Five factor structure. In Personality and personality disorders (Routledge), pp. 34–47.
2. McCrae, R.R., and Costa, P.T. (2008). Empirical and theoretical status of the five-factor model of personality traits. In The SAGE handbook of personality theory and assessment (SAGE Publications Ltd), pp. 273–294.
3. McCrae, R.R., Costa, P.T., Jr., Del Pilar, G.H., Rolland, J.-P., and Parker, W.D. (1998). Cross-cultural assessment of the five-factor model: The Revised NEO Personality Inventory. J. Cross Cult. Psychol. *29*, 171–188. https://doi.org/10.1177/0022022198291009.
4. Galton, F. (1884). The measurement of character. Fortn. Rev. *42*, 179–185.
5. Allport, G.W., and Odbert, H.S. (1936). Trait-names: A psycho-lexical study. Psychol. Monogr. *47*, i-171. https://doi.org/10.1037/h0093360.
6. Cattell, R.B. (1943). The description of personality: Basic traits resolved into clusters. J. Abnorm. Soc. Psychol. *38*, 476–506. https://doi.org/10.1037/h0054116.
7. Cattell, R.B. (1945). The description of personality: Principles and findings in a factor analysis. Am. J. Psychol. *58*, 69–90. https://doi.org/10.2307/1417576.
8. Fiske, D.W. (1949). Consistency of the factorial structures of personality ratings from different sources. J. Abnorm. Soc. Psychol. *44*, 329–344. https://doi.org/10.1037/h0057198.
9. Tupes, E.C., and Christal, R.E. (1992). Recurrent personality factors based on trait ratings. J. Pers. *60*, 225–251. https://doi.org/10.1111/j.1467-6494.1992.tb00973.x.
10. Goldberg, L.R. (1981). Language and individual differences: The search for universals in personality lexicons. Rev. Pers. Soc. Psychol. *2*, 141–165.
11. Peabody, D., and Goldberg, L.R. (1989). Some determinants of factor structures from personality-trait descriptors. J. Pers. Soc. Psychol. *57*, 552–567. https://doi.org/10.1037//0022-3514.57.3.552.
12. Digman, J.M., and Takemoto-Chock, N.K. (1981). Factors in the natural language of personality: Re-analysis, comparison, and interpretation of six major studies. Multivariate Behav. Res. *16*, 149–170. https://doi.org/10.1207/s15327906mbr1602_2.
13. McCrae, R.R., and Costa, P.T. (1987). Validation of the five-factor model of personality across instruments and observers. J. Pers. Soc. Psychol. *52*, 81–90. https://doi.org/10.1037//0022-3514.52.1.81.
14. McCrae, R.R., and John, O.P. (1992). An introduction to the five-factor model and its applications. J. Pers. *60*, 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x.
15. Roccas, S., Sagiv, L., Schwartz, S.H., and Knafo, A. (2002). The Big Five personality factors and personal values. Pers. Soc. Psychol. Bull. *28*, 789–801. https://doi.org/10.1177/0146167202289008.
16. Azucar, D., Marengo, D., and Settanni, M. (2018). Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. Pers. Individ. Dif. *124*, 150–159. https://doi.org/10.1016/J.PAID.2017.12.018.
17. Majumder, N., Poria, S., Gelbukh, A., and Cambria, E. (2017). Deep Learning-Based Document Modeling for Personality Detection from Text. IEEE Intell. Syst. *32*, 74–79. https://doi.org/10.1109/MIS.2017.23.
18. Ren, Z., Shen, Q., Diao, X., and Xu, H. (2021). A sentiment-aware deep learning approach for personality detection from text. Inf. Process. Manag. *58*, 102532. https://doi.org/10.1016/J.IPM.2021.102532.
19. Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., and Mian, A. (2023). A comprehensive overview of large language models. ACM Trans. Intell. Syst. Technol. *16*, 1–72.
20. Fan, L., Li, L., Ma, Z., Lee, S., Yu, H., and Hemphill, L. (2024). A bibliometric review of large language models research from 2017 to 2023. ACM Trans. Intell. Syst. Technol. *15*, 1–25. https://doi.org/10.1145/3664930.
21. Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., and Dong, Z. (2026). A Survey of Large Language Models. Front. Comput. Sci. *20*, 2012627. https://doi.org/10.1007/S11704-026-60308-3.

22. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog *1*, 9.

23. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., and Askell, A. (2020). Language models are few-shot learners. Adv. Neural Inf. Process. Syst. *33*, 1877–1901.

24. Larochelle, H., Erhan, D., and Bengio, Y. (2008). Zero-data learning of new tasks. In AAAI, p. 3.

25. Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S.A., Feder, A., Emanuel, D., Cohen, A., et al. (2022). Shared computational principles for language processing in humans and deep language models. Nat. Neurosci. *25*, 369–380. https://doi.org/10.1038/s41593-022-01026-4.

26. Cronbach, L.J. (1951). Coefficient Alpha and the Internal Structure of Tests. Psychometrika *16*, 297–334. https://doi.org/10.1007/BF02310555.

27. Nunnally, J.C., and Bernstein, I.H. (1994). Psychometric theory (McGraw Hill).

28. Fabrigar, L.R., and Wegener, D.T. (2012). Exploratory Factor Analysis (Oxford University Press).

29. Brown, T.A. (2015). Confirmatory Factor Analysis for Applied Research (Guilford Publications).

30. Ozer, D.J., and Benet-Martínez, V. (2006). Personality and the prediction of consequential outcomes. Annu. Rev. Psychol. *57*, 401–421. https://doi.org/10.1146/annurev.psych.57.102904.190127.

31. Roberts, B.W., Kuncel, N.R., Shiner, R., Caspi, A., and Goldberg, L.R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. Perspect. Psychol. Sci. *2*, 313–345. https://doi.org/10.1111/j.1745-6916.2007.00047.x.

32. Cronbach, L.J., and Meehl, P.E. (1955). Construct validity in psychological tests. Psychol. Bull. *52*, 281–302. https://doi.org/10.1037/H0040957.

33. DeVellis, R.F., and Thorpe, C.T. (2021). Scale Development: Theory and Applications (Sage Publications).

34. Hommel, B.E., Wollang, F.-J.M., Kotova, V., Zacher, H., and Schmukle, S.C. (2022). Transformer-based deep neural language modeling for construct-specific automatic item generation. Psychometrika *87*, 749–772. https://doi.org/10.1007/s11336-021-09823-9.

35. Götz, F.M., Maertens, R., Loomba, S., and van der Linden, S. (2023). Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development. Psychol. Methods *29*, 494–518.

36. Hernandez, I., and Nie, W. (2023). The AI-IP: Minimizing the guesswork of personality scale item development through artificial intelligence. Pers. Psychol. *76*, 1011–1035. https://doi.org/10.1111/peps.12543.

37. Lee, P., Fyffe, S., Son, M., Jia, Z., and Yao, Z. (2023). A paradigm shift from "human writing" to "machine generation" in personality test development: An application of state-of-the-art natural language processing. J. Bus. Psychol. *38*, 163–190. https://doi.org/10.1007/s10869-022-09864-6.

38. American Psychiatric Association (2013). Diagnostic and Statistical Manual of Mental Disorders, 5th ed. (American Psychiatric Association).

39. Woolfolk, J.M. (2012). The Only Astrology Book You'll Ever Need: Now with an Interactive PC-And Mac-Compatible CD (Taylor Trade Publications).

40. John, O.P., Donahue, E.M., and Kentle, R.L. (1991). The Big Five Inventory—Versions 4a and 54 (University of California).

41. OpenAI (2023). GPT-4 Technical Report (OpenAI).

42. Bosch User Manual for Model HBT578FS2A [PDF Manual]. Appliances Online. https://commercial.appliancesonline.com.au/public/manuals/HBT578FS2A-Bosch-User-Manual.pdf.

43. Tolkien, J.R.R. (1954). The Lord of the Rings (George Allen & Unwin).

44. Marsh, H.W., Hau, K.T., and Wen, Z. (2004). In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's (1999) Findings. Struct. Equ. Model.: A Multidiscip. J. *11*, 320–341. https://doi.org/10.1207/S15328007SEM1103_2.

45. Ke, L., Tong, S., Cheng, P., and Peng, K. (2025). Exploring the Frontiers of Llms in Psychological Applications: A Comprehensive Review. Artif. Intell. Rev. *58*, 305. https://doi.org/10.1007/S10462-025-11297-5.

46. Guo, Z., Lai, A., Thygesen, J.H., Farrington, J., Keen, T., and Li, K. (2024). Large Language Models for Mental Health Applications: Systematic Review. JMIR Ment. Health *11*, e57400. https://doi.org/10.2196/57400.

47. Coughlan, G., Coutrot, A., Khondoker, M., Minihane, A.-M., Spiers, H., and Hornberger, M. (2019). Toward personalized cognitive diagnostics of at-genetic-risk Alzheimer's disease. Proc. Natl. Acad. Sci. USA *116*, 9285–9292. https://doi.org/10.1073/pnas.1901600116.

48. Kunz, L., Schröder, T.N., Lee, H., Montag, C., Lachmann, B., Sariyska, R., Reuter, M., Stirnberg, R., Stöcker, T., Messing-Floeter, P.C., et al. (2015). Reduced grid-cell–like representations in adults at genetic risk for Alzheimer's disease. Science *350*, 430–433. https://doi.org/10.1126/science.aac8128.

49. Bierbrauer, A., Kunz, L., Gomes, C.A., Luhmann, M., Deuker, L., Getzmann, S., Wascher, E., Gajewski, P.D., Hengstler, J.G., Fernandez-Alvarez, M., et al. (2020). Unmasking selective path integration deficits in Alzheimer's disease risk carriers. Sci. Adv. *6*, eaba1394. https://doi.org/10.1126/sciadv.aba1394.

50. Teipel, S., Singh, D., Dubbelman, M.A., Pan, G., Tang, Y., and König, A. (2025). Early Detection of Alzheimer's Disease: From Multiplex Assays and Imaging to Point of Care Devices and AI-Based Functional Monitoring (London, England: SAGE Publications Sage UK).

51. Lho, S.K., Park, S.C., Lee, H., Oh, D.Y., Kim, H., Jang, S., Jung, H.Y., Yoo, S.Y., Park, S.M., and Lee, J.Y. (2025). Large Language Models and Text Embeddings for Detecting Depression and Suicide in Patient Narratives. JAMA Netw. Open *8*, e2511922. https://doi.org/10.1001/JAMANETWORKOPEN.2025.11922.

52. Kim, J., Ma, S.P., Chen, M.L., Galatzer-Levy, I.R., Torous, J., van Roessel, P.J., Sharp, C., Pfeffer, M.A., Rodriguez, C.I., Linos, E., and Chen, J.H. (2025). Optimizing large language models for detecting symptoms of depression/anxiety in chronic diseases patient communications. npj Digit. Med. *8*, 580. https://doi.org/10.1038/s41746-025-01969-5.

53. Shin, D., Kim, H., Lee, S., Cho, Y., and Jung, W. (2024). Using Large Language Models to Detect Depression From User-Generated Diary Text Data as a Novel Approach in Digital Mental Health Screening: Instrument Validation Study. J. Med. Internet Res. *26*, e54617. https://doi.org/10.2196/54617.

54. Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E., and Tatham, R.L. (2009). Multivariate data analysis. Cranbury (Pearson Education).

55. Robinson, J.P., Shaver, P.R., and Wrightsman, L.S. (2013). Measures of Personality and Social Psychological Attitudes: Measures of Social Psychological Attitudes (Academic Press).

56. Hopwood, C.J., and Donnellan, M.B. (2010). How should the internal structure of personality inventories be evaluated? Pers. Soc. Psychol. Rev. *14*, 332–346. https://doi.org/10.1177/1088868310361240.

57. Argyle, L.P., Busby, E.C., Fulda, N., Gubler, J.R., Rytting, C., and Wingate, D. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. Polit. Anal. *31*, 337–351. https://doi.org/10.1017/PAN.2023.2.

58. Park, J.S., O'Brien, J., Cai, C.J., Morris, M.R., Liang, P., and Bernstein, M.S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. UIST 2023 - Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. https://doi.org/10.1145/3586183.3606763;TOPIC:TOPIC:CONFERENCE-COLLECTIONS.

59. Aher, G.V., Arriaga, R.I., and Kalai, A.T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. In International conference on machine learning (PMLR), pp. 337–371.

60. Kumar, S., Sumers, T.R., Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K.A., Griffiths, T.L., Hawkins, R.D., and Nastase, S.A. (2024). Shared

functional specialization in transformer-based language models and the human brain. Nat. Commun. *15*, 5523. https://doi.org/10.1038/s41467-024-49173-5.

61. Michelmann, S., Kumar, M., Norman, K.A., and Toneva, M. (2025). Large language models can segment narrative events similarly to humans. Behav. Res. Methods *57*, 39. https://doi.org/10.3758/s13428-024-02569-z.
62. Saucier, G., and Srivastava, S. (2015). What Makes a Good Structural Model of Personality? Evaluating the Big Five and Alternatives (American Psychological Association).
63. Stewart, R.D., Mõttus, R., Seeboth, A., Soto, C.J., and Johnson, W. (2022). The finer details? The predictability of life outcomes from Big Five domains, facets, and nuances. J. Pers. *90*, 167–182. https://doi.org/10.1111/jopy.12660.
64. Tao, Y., Viberg, O., Baker, R.S., and Kizilcec, R.F. (2024). Cultural bias and cultural alignment of large language models. PNAS Nexus *3*, pgae346. https://doi.org/10.1093/pnasnexus/pgae346.
65. Gallegos, I.O., Rossi, R.A., Barrow, J., Tanjim, M.M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., and Ahmed, N.K. (2024). Bias and fairness in large language models: A survey. Comput. Linguist. *50*, 1097–1179. https://doi.org/10.1162/coli_a_00524.
66. Joinson, A. (1999). Social desirability, anonymity, and Internet-based questionnaires. Behav. Res. Methods Instrum. Comput. *31*, 433–438. https://doi.org/10.3758/bf03200723.
67. Kreuter, F., Presser, S., and Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity. Public Opin. Q. *72*, 847–865. https://doi.org/10.1093/poq/nfn063.
68. Peer, E., Brandimarte, L., Samat, S., and Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. J. Exp. Soc. Psychol. *70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006.
69. Peer, E., Rothschild, D., Gordon, A., Evernden, Z., and Damer, E. (2022). Data quality of platforms and panels for online behavioral research. Behav. Res. Methods *54*, 1643–1662. https://doi.org/10.3758/s13428-021-01694-3.
70. Atari, M., Xue, M., Park, P., Blasi, D., and Henrich, J. (2023). Which Humans? Preprint at PsyArXiv. https://doi.org/10.31234/osf.io/5b26t.
71. Roberts, B.W., Walton, K.E., and Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies. Psychol. Bull. *132*, 1–25. https://doi.org/10.1037/0033-2909.132.1.1.
72. Palan, S., and Schitter, C. (2018). Prolific. ac—A subject pool for online experiments. J. Behav. Exp. Finance *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004.
73. Ferrando, P.J. (2008). The impact of social desirability bias on the EPQ-R item scores: An item response theory analysis. Pers. Individ. Dif. *44*, 1784–1794. https://doi.org/10.1016/j.paid.2008.02.005.
74. Cronbach, L.J. (1942). An analysis of techniques for diagnostic vocabulary testing. J. Educ. Res. *36*, 206–217. https://doi.org/10.1080/00220671.1942.10881160.
75. Paulhus, D.L. (1991). Measurement and control of response bias. In Measures of Personality and Social Psychological Attitudes, J.P. Robinson, P.R. Shaver, and L.S. Wrightsman, eds. (Elsevier), pp. 17–59. https://doi.org/10.1016/B978-0-12-590241-0.50006-X.
76. Podsakoff, P.M., MacKenzie, S.B., Lee, J.-Y., and Podsakoff, N.P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. J. Appl. Psychol. *88*, 879–903. https://doi.org/10.1037/0021-9010.88.5.879.
77. Warstadt, A., Singh, A., and Bowman, S.R. (2019). Neural network acceptability judgments. Trans. Assoc. Comput. Linguist. *7*, 625–641. https://doi.org/10.1162/tacl_a_00290.
78. Shrout, P.E., and Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychol. Bull. *86*, 420–428. https://doi.org/10.1037/0033-2909.86.2.420.
79. Koo, T.K., Li, M.Y., Koo, T.K., and Li, M.Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J. Chiropr. Med. *15*, 155–163. https://doi.org/10.1016/J.JCM.2016.02.012.
80. Meade, A.W., and Craig, S.B. (2012). Identifying careless responses in survey data. Psychol. Methods *17*, 437–455. https://doi.org/10.1037/a0028085.
81. Huebner, E.S. (1994). Preliminary development and validation of a multidimensional life satisfaction scale for children. Psychol. Assess. *6*, 149–158. https://doi.org/10.1037/1040-3590.6.2.149.
82. Kroenke, K., Spitzer, R.L., Williams, J.B.W., and Löwe, B. (2009). An ultra-brief screening scale for anxiety and depression: the PHQ–4. Psychosomatics *50*, 613–621. https://doi.org/10.1176/appi.psy.50.6.613.
83. Martinussen, L.M., Lajunen, T., Møller, M., and Özkan, T. (2013). Short and user-friendly: The development and validation of the Mini-DBQ. Accid. Anal. Prev. *50*, 1259–1265. https://doi.org/10.1016/j.aap.2012.09.030.
84. Kang, W., and Malvaso, A. (2024). Understanding the longitudinal associations between e-cigarette use and general mental health, social dysfunction and anhedonia, depression and anxiety, and loss of confidence in a sample from the UK: A linear mixed effect examination. J. Affect. Disord. *346*, 200–205. https://doi.org/10.1016/j.jad.2023.11.013.
85. Timmerman, M.E., and Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. Psychol. Methods *16*, 209–220. https://doi.org/10.1037/A0023353.
86. Browne, M.W. (2000). Cross-Validation Methods. J. Math. Psychol. *44*, 108–132. https://doi.org/10.1006/JMPS.1999.1279.
87. Yarkoni, T., and Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. Perspect. Psychol. Sci. *12*, 1100–1122. https://doi.org/10.1177/1745691617693393.
88. Waldmann, P., Mészáros, G., Gredler, B., Fuerst, C., and Sölkner, J. (2013). Evaluation of the lasso and the elastic net in genome-wide association studies. Front. Genet. *4*, 270. https://doi.org/10.3389/fgene.2013.00270.
89. Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B: Stat. Methodol. *67*, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.
90. Polat, G., Ergenc, I., Kani, H.T., Alahdab, Y.O., Atug, O., and Temizel, A. (2022). Class distance weighted cross-entropy loss for ulcerative colitis severity estimation. In Annual Conference on Medical Image Understanding and Analysis (Springer), pp. 157–171.
91. Savcisens, G., Eliassi-Rad, T., Hansen, L.K., Mortensen, L.H., Lilleholt, L., Rogers, A., Zettler, I., and Lehmann, S. (2024). Using sequences of life-events to predict human lives. Nat. Comput. Sci. *4*, 43–56. https://doi.org/10.1038/s43588-023-00573-5.

Please cite this article in press as: Monsa et al., Generating and analyzing personality questionnaires using large language models, iScience (2026), https://doi.org/10.1016/j.isci.2026.116909

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Behavioral data and LLM prompts	This paper	OSF: https://osf.io/3t8zh/overview?view_only=547e7d3775c545e5aea6174c4845cc7b
Software and algorithms		
Python version 3.9.5	Python Software Foundation	https://www.python.org
R version 4.0.3	R software	https://cran.r-project.org
GPT-4 (OpenAI API)	OpenAI	https://platform.openai.com
Prolific	Prolific	https://www.prolific.com
QuestionPro	QuestionPro Inc.	https://www.questionpro.com
Data analysis code	This paper	https://osf.io/3t8zh/overview?view_only=547e7d3775c545e5aea6174c4845cc7b
Other		
DSM-5 personality disorders descriptions	American Psychiatric Association (2013). Diagnostic and statistical manual of mental disorders (5th ed.)[38]	https://psycnet.apa.org/record/2013-14907-000
Astrology text excerpts	Woolfolk et al.[39]	Copyrighted book source: The Only Astrology Book You'll Ever Need
Literary text excerpts	Tolkien[43]	Copyrighted book source: The Lord of the Rings
User manual text	Bosch[42]	https://commercial.appliancesonline.com.au/public/manuals/HBT578FS2A-Bosch-User-Manual.pdf

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Participants

Participants (N = 600) were recruited through Prolific Academic (https://www.prolific.com), an online platform that connects researchers with diverse, pre-screened participants.[72] Eligibility criteria included being at least 18 years old, fluent in English, and residing in either the United States or the United Kingdom. The study was approved by the School of Computer Science and Engineering Committee for the Use of Human Subjects in Research, The Hebrew University of Jerusalem (approval no.: CSE-2024-05), and all participants provided informed consent. Participants completed the survey through QuestionPro (www.questionpro.com) and were compensated through Prolific in accordance with platform guidelines.

A total of 10 participants failed embedded attention checks and were excluded. Additionally, participants with implausibly fast completion times (less than 2 s per item) were excluded (n = 2), leaving a final sample of 588 participants (324 females, 264 males; mean age = 40.05 years, SD = 12.69). To assess whether exclusions introduced systematic bias, excluded participants were compared to the retained sample on all demographic variables. No significant differences were found (all p-values >0.05). The sample showed demographic variability across marital status, education, employment, and household income (see Table S12).

METHOD DETAILS

Personality item generation

Personality item generation using GPT-4

We developed an automated procedure to generate personality items using OpenAI's GPT-4 model.[41] We designed a prompt to guide GPT-4 in generating relevant questionnaire items for a variety of conceptual personality models using a few-shot learning approach[23] (Figure 1). Few-shot learning was implemented by providing the model with six examples from the validated Big Five Inventory (BFI-44[40]) to ensure the generated questions follow the same format. The BFI is a widely used tool in surveys and research for assessing personality based on the five factors of the FFM. The BFI examples were selected to represent both direct and reverse keyed items across all five factors of the FFM, where reverse questions require responses in the opposite direction of the trait being assessed. Items were generated via the OpenAI API using default model parameters (temperature = 1.0), with a single API call per construct.

Please cite this article in press as: Monsa et al., Generating and analyzing personality questionnaires using large language models, iScience (2026), https://doi.org/10.1016/j.isci.2026.116909

The prompt was structured to ensure adaptability across different texts while maintaining consistency in the quality of the generated questions. Items were constrained to 10 words to ensure clarity and minimize cognitive load, and the model was instructed to generate items that would elicit responses across the full range of the Likert scale to maximize response variability. To minimize response bias, the prompt required generation of both positively keyed (direct) and negatively keyed (reverse) items, that is, statements phrased in both direct and reverse directions (e.g., "I enjoy social gatherings" vs. "I avoid social interaction") to reduce the tendency to agree with items regardless of content.[55,73–75] Items were also required to capture the key traits described in the source texts and reflect the core dimensions of each personality construct.

By structuring the prompt with clear instructions for processing source texts and specifying the desired item characteristics, GPT-4 was able to generate items that were not only consistent within the source texts and designed to elicit varied responses. The model generated five questionnaire items per construct, intended to reflect the associated personality traits.

Source materials

To test our process for personality item generation, we selected two distinct text sources describing personality: the clinical descriptions of personality disorders from the Diagnostic and Statistical Manual of Mental Disorders (DSM-5)[38] and personality descriptions of zodiac signs from "The Only Astrology Book You'll Ever Need".[39] As described in the Introduction, two textual sources were chosen to represent different ends of the spectrum - one being scientifically established clinical classification system describing personality disorders, and the other being popular but scientifically unsupported character typologies. This contrast allows us to evaluate the LLM's ability to generate coherent personality items regardless of the specific content of the source material.

For the DSM-based questionnaire, texts were drawn from the Personality Disorders chapter in Section II (Diagnostic Criteria and Codes) of the DSM-5 (pages 646–681). Specifically, for each disorder, we extracted the "Diagnostic Features" and "Associated Features Supporting Diagnosis" subsections, as these provide narrative descriptions of how each disorder manifests behaviorally, and are more informative for item generation than the structured diagnostic criteria used for clinical classification. Input texts ranged from 561 to 1,321 words per disorder (mean = 895, SD = 224 words; Table S13).

For the astrology-based questionnaire, texts were drawn from Part 1 (Sun Sign Astrology), Section 1 (Sun Signs) of *The Only Astrology Book You'll Ever Need* (pages 9–68). For each zodiac sign, we extracted the sections describing personality traits, selecting passages comparable in style and content to the DSM narrative descriptions to ensure consistency across sources. Input texts ranged from 978 to 1,561 words per sign (mean = 1,268, SD = 164 words; Table S13).

This approach enabled us to generate 50 items for the 10 DSM-5 disorders (facets; 5 items per disorder) and 60 items for the 12 zodiac signs (facets; 5 items per sign). For analysis, DSM-based items were aggregated into their standard higher-order clusters (Clusters A, B, and C), which we treated as the model's factors. Likewise, astrology items were aggregated into the four classical elements (Water, Fire, Air, Earth), which served as the factors for the astrology model.

To evaluate the extent to which questionnaire content is influenced by the input text versus the language model itself, we applied our previously developed GPT-4-based few-shot questionnaire generation procedure to two texts unrelated to personality: (1) a technical manual for a Bosch built-in oven user Manual,[42] and (2) a literary landscape description from *The Lord of the Rings*.[43] The generation prompts included example BFI items as before, but the source texts provided no explicit trait information.

LLM-based item validation

Item quality assessment with GPT-4

To assess the psychometric quality of the generated personality items, we used GPT-4 as an expert reviewer by prompting it to evaluate each item according to standard psychometric criteria. The model was given a structured prompt positioning it as a psychometrics expert and instructed to assess each item on three dimensions: (1) identification of potential issues (e.g., ambiguity, double-barreledness, extreme wording, unclear valence, or susceptibility to social desirability bias),[33,76] (2) suggestions for improvement, and (3) an overall quality score ranging from 1 (very poor) to 5 (excellent). Each item was inserted into the prompt and the model's structured response was recorded. This procedure allowed us to estimate item-level quality systematically using the LLM's internal representation of psychometric standards, providing an additional layer of content validation prior to empirical testing.

In addition to the GPT-based evaluation, all items were independently reviewed by the research team, who verified the LLM's assessments using the same criteria (ambiguity, extreme wording, social desirability bias). This manual review served as an independent confirmation layer to mitigate potential problems arising from exclusive reliance on the LLM.

Each item was assessed independently in three separate GPT-4 runs and scores were averaged across runs. Items with a mean score below 3, the scale midpoint, were flagged for revision. In the DSM-based questionnaire, out of 50 items, 30 items received a score of 3 or higher, indicating high clarity and Likert suitability. 20 of the items were flagged for potential issues, most commonly due to double-barreled phrasing (e.g., combining multiple behaviors into one item), emotionally charged or extreme wording, or overly complex sentence structure. Based on these reviews, 10 items were revised using GPT-4, And 2 items were revised manually. The remaining flagged items were retained without revision, because the concerns were judged to be minor by the research team. The same quality assessment procedure was applied to the astrology-based questionnaire. Of 60 items, 20 were flagged for potential issues. However, none were revised following manual review, as the flagged issues were judged to reflect the descriptive language of the source astrological texts rather than correctable psychometric flaws.

Please cite this article in press as: Monsa et al., Generating and analyzing personality questionnaires using large language models, iScience (2026), https://doi.org/10.1016/j.isci.2026.116909

Content validity with GPT-4-based factor discrimination

To validate the content of generated questionnaire items, we used GPT-4 for content validation of the generated items. Each item was compared against the personality descriptions of all other facets (e.g., personality disorders for the DSM-based questionnaire). Specifically, for every item, GPT-4 was given two trait descriptions at a time: Text A, the description of the facet from which the item was generated, and Text B, the description of another facet. GPT-4 was prompted as follows: *"You are given two personality trait descriptions (Text A and Text B) and one questionnaire item. Your task is to decide which description the item aligns with more closely. If the item ends with "(R)", it is reverse-scored and reflects the opposite of the trait described"* (for full prompt, see Figure S1). This procedure was repeated for each item against every other facet's description.

A net alignment score was computed as the sum of pairwise comparison scores (aligned = +1, misaligned = −1, unclear = 0), ranging from −9 to 9. Items with a net score below 4 were excluded to ensure sufficient construct specificity and replaced by prompting GPT-4 to generate a new item conveying the same construct. Scores across items were heavily skewed toward high alignment with 94% of items scoring 7 or above and no items scoring between 4 and 7, indicating a natural gap in the distribution. Items scoring below this threshold represented clear outliers and were considered to lack sufficient content specificity. Because unclear judgments (0) do not affect the score, the threshold reflects only the balance between aligned and misaligned judgments. For example, the item "Always seeks admiration" (Narcissistic PD) received a perfect score of 9, indicating clear construct specificity, whereas the item "Doesn't care about others' opinions" (Avoidant PD, reverse-scored) received a score of 3 due to multiple misaligned judgments, reflecting insufficient specificity, and was subsequently replaced. The questionnaire-level alignment score is derived by averaging item-level net scores and dividing by 9, yielding a normalized score ranging from −1 to 1 (Figure S5).

The alignment scoring procedure was applied only to the DSM-based questionnaire, as it requires theoretically distinct named constructs against which item discrimination can be formally evaluated. The elemental groupings of the astrology-based questionnaire (Fire, Earth, Air, Water) do not meet this criterion, as they do not define clearly separable constructs, rendering pairwise discrimination uninformative.

For the DSM-based questionnaire 38 out of 50 had a perfect average score of 1, and one item was filtered out and replaced by prompting GPT-4 to generate replacement items conveying the same concept. The total normalized score of the DSM-based questionnaire was 0.94. After revision, 40 out of 50 items had a perfect average score of 1, and the total score was 0.96.

Structural and linguistic analysis of questionnaire items

We assessed structural and linguistic properties of the final questionnaire items by computing item length (number of words per item), readability using the Flesch Reading Ease score, and grammatical acceptability using a RoBERTa-base classifier fine-tuned on the Corpus of Linguistic Acceptability (CoLA[36,77]). These metrics were computed for DSM-based and astrology-based items, as well as for BFI items for comparison.

Correlation estimation

We assessed the structural coherence of the generated items by estimating inter-item correlations using GPT-4. The model was prompted to provide Pearson correlation estimates for item pairs, with two types of pairings: within-factor (all possible item pairs within a facet) and between-facet (each item compared with one randomly selected item from each of the other facets) (full prompt shown in Figure S2). To calibrate GPT-4's judgments, each prompt included ten reference pairs of IPIP items from the Eugene-Springfield Community Sample, spanning five correlation levels: high negative (≤ -0.5), moderate negative (-0.5 to -0.2), neutral (-0.2 to 0.2), moderate positive (0.2–0.5), and high positive (≥ 0.5). For each target pair, GPT-4 returned an estimated correlation value between −1 and +1. These estimates were then compared to actual item correlations after collecting responses from 600 participants using mean absolute error (MAE) and Pearson correlation between predicted and observed values. To quantify the precision of the correlation estimates, we computed 95% confidence intervals using 1,000 bootstrap resamples across item pairs.

Average response estimation

We assessed GPT-4's ability to predict response distributions by having the model estimate the expected mean response for each generated item on a 1–5 Likert scale. The model was prompted to provide these estimates assuming presentation to a large, diverse adult population, without providing reference examples (full prompt shown in Figure S3). These estimates were then validated against actual mean item responses from the same 600-participant sample, using mean absolute error (MAE) and Pearson correlation between predicted and observed values as accuracy metrics. Confidence intervals for the correlation between predicted and observed mean responses were estimated using 1,000 bootstrap resamples of item-level pairs. As the absolute level of responses is directly meaningful in this context, we additionally computed intraclass correlation (ICC2; two-way random effects model, single raters[78,79]) as a measure of absolute agreement between GPT-4 estimated and empirical mean responses.

Human participants study

Personality assessments and outcome measures

Each participant completed all three personality assessments, allowing for direct within-subject comparisons across models: the Big Five Inventory (BFI-44),[40] a DSM-based questionnaire, and an astrology-based questionnaire. The latter two were generated

using the GPT-4-based procedure described earlier. Each questionnaire was presented on a separate page in randomized order. Participants were blind to the origin of each questionnaire and were not informed that any items had been generated by a language model. All items were presented in a 5-point Likert format and interspersed with attention-check questions to ensure response quality.[80] To assess external validity, participants completed a battery of life outcome measures (adapted from Stewart et al.[63]), including subjective well-being (Life Satisfaction Scale),[81] anxiety and depression symptoms (PHQ-4),[82] risky driving behaviors,[83] as well as demographic and behavioral variables such as political orientation, marital status, household income, education level, alcohol and tobacco use and employment status. Tobacco and alcohol use were included as behavioral outcome measures, consistent with evidence linking substance use to personality traits and mental health.[84] These measures provided a broad array of real-world outcomes to evaluate the predictive validity of each personality model (See Table S14 for full life outcomes questionnaires).

Confirmatory and exploratory factor analyses (CFA and EFA)

To evaluate the internal structure of the questionnaires, we employed both exploratory and confirmatory factor analytic approaches within a unified validation framework. Confirmatory factor analysis (CFA) was conducted to test the correspondence between observed data and the theoretical grouping structures of each questionnaire. For each questionnaire, a theory-driven measurement model was specified: five factors for the BFI, three DSM-based clusters, and four classical elements for the astrology-based questionnaire. Each item was assigned to load on a single predefined latent factor, with no cross-loadings estimated, and latent factors were allowed to correlate. Model parameters were estimated using maximum likelihood estimation. Model identification was achieved by fixing one loading per latent factor to 1.0. Model fit was evaluated using χ^2, CFI, TLI and RMSEA. Consistent with psychometric recommendations, CFA results were interpreted as indicators of structural correspondence rather than strict pass/fail criteria, given the known sensitivity of CFA fit indices in broad personality inventories.[44,56] Subsequently, exploratory factor analysis (EFA) was conducted to examine the empirical structure of item responses without imposing theoretical constraints. Prior to EFA, data suitability was assessed using the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and Bartlett's test of sphericity. KMO values indicated good sampling adequacy for all questionnaires (BFI = 0.89; DSM-based = 0.88; astrology-based = 0.89), and Bartlett's test was significant in all cases (p-value <0.001), indicating sufficient correlations among items for factor analysis. EFA was performed using principal axis factoring with oblimin rotation to allow correlated factors. The number of factors was determined using permutation-based parallel analysis (100 permutations, 95th percentile criterion[85]), appropriate for ordered polytomous items. Loadings below |0.30| were not interpreted.

Predictive modeling of life outcomes

We aimed to evaluate the predictive power of each personality model for a range of real-world life outcomes. **Predictors:** For each participant, the predictors consisted of responses to one of three personality questionnaires: the BFI-44 (44 items), a DSM-based questionnaire (50 items), or an astrology-based questionnaire (60 items). Each questionnaire included a distinct set of items answered on a 5-point Likert scale (1–5). **Outcomes:** Ordinal outcomes comprised four questions from the subjective well-being questionnaire, four questions from the anxiety and depression questionnaire, nine questions from the risky driving questionnaire, as well as household income, education level, and alcohol and tobacco use. Non-ordinal categorical outcomes included employment status, marital status, and political orientation.

To estimate predictive performance while minimizing overfitting, we applied 10-fold cross-validation, a standard approach for quantifying out-of-sample prediction accuracy.[86] The dataset was randomly partitioned into 10 equal-sized parts (folds), with models trained on 9-folds and tested on the remaining withheld fold, repeating this procedure for all folds and averaging performance. This approach provides an unbiased estimate of generalization performance, which is particularly important in personality-to-outcome prediction where predictors are correlated and effect sizes are modest.[87]

Three personality questionnaires were considered: the BFI-44, the DSM-based, and the astrology-based. Each questionnaire consisted of Likert-scale (1–5) participant responses to a distinct set of questionnaire items. Participants' responses were mean-centered but not standardized to preserve the ordinal nature of the scales. Predictors were standardized, and age and gender were included as covariates in all models, with age entered both linearly and quadratically.

For ordinal outcomes, we applied logistic ordinal regression with elastic net regularization,[63,88,89] modeling each outcome independently. For non-ordinal categorical outcomes, including employment status, marital status, and political orientation, we used multinomial logistic regression with elastic net regularization within the same 10-fold cross-validation framework. These outcomes were nominal rather than ordinal and therefore required a different regression approach.

Model performance was assessed using quadratic weighted kappa (QWK) for ordinal outcomes and Cohen's kappa for non-ordinal categorical outcomes. Both metrics were computed per fold and per outcome. QWK was selected for ordinal outcomes because it accounts for the magnitude of disagreements, assigning greater penalties to larger discrepancies.[90,91] Cohen's kappa was used for nominal outcomes as it captures agreement beyond chance without assuming an order.

For each outcome, we statistically compared the models by first performing a Friedman test across the fold-wise kappa scores to assess overall model differences. When the Friedman test was significant (p-value <0.05), we followed up with pairwise Wilcoxon signed-rank tests on the fold-paired scores and applied false discovery rate (FDR) correction for multiple comparisons. The same procedure was used for both ordinal and nominal outcomes.

Please cite this article in press as: Monsa et al., Generating and analyzing personality questionnaires using large language models, iScience (2026), https://doi.org/10.1016/j.isci.2026.116909

In addition to item-level analyses, we conducted a complementary scale-level analysis in which factor scores were used as predictors. Factor scores were computed as the average response across items within each factor. For the DSM-based questionnaire, factors corresponded to the three DSM personality disorder clusters (Clusters A, B, C). For the BFI, factors corresponded to the five standard Big Five traits. For the astrology-based questionnaire, factors corresponded to the four classical elements (Fire, Earth, Air, Water). The same 10-fold cross-validation framework, outcome measures, and statistical comparison procedure described above were applied to the factor-score models.

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical details are reported in the results section, figure legends, and Tables S1–S14 $n = 588$ participants (each representing one independent observation; exclusion criteria are described in Experimental Model and Study Participant Details). Center is reported as mean; dispersion is reported as SD or 95% bootstrap confidence intervals (1,000 resamples) as indicated. Statistical significance was defined as FDR-corrected $p < 0.05$. Statistical tests used include Cronbach's alpha (internal consistency; Tables S4 and S5), confirmatory factor analysis (χ^2, CFI, TLI, RMSEA; Tables S6–S8), exploratory factor analysis with permutation-based parallel analysis, Pearson correlation and ICC2 (GPT-4 estimation accuracy; Figures 3 and 4), and Friedman tests with post-hoc Wilcoxon signed-rank tests and FDR correction (model comparisons; Tables S9–S11). Predictive performance was evaluated with 10-fold cross-validation and reported as quadratic weighted kappa (ordinal outcomes) or Cohen's kappa (categorical outcomes). All analyses were performed in Python and R.