LLM-Generated Social-Desirability Ratings for Forced-Choice Questionnaire Construction: A Preregistered Downstream Validation

Yui Furukawa
The University of Tokyo
yui-furukawa@p.u-tokyo.ac.jp

Kyosuke Bunji
Kobe University
bunji@bear.kobe-u.ac.jp

Kensuke Okada
The University of Tokyo
ken@p.u-tokyo.ac.jp

Abstract

Large language models (LLMs) increasingly produce evaluative judgments, but rating-level human–model agreement does not establish whether they can replace human inputs in a downstream task. Using HEXACO personality items as a case study, we conduct a preregistered three-stage validation of whether LLM-generated social-desirability ratings can serve as inputs to forced-choice scale construction, a psychometric task that requires costly human ratings. In Stage I, three LLMs served as raters of item-level social desirability for sentence- and adjective-format items, with human ratings serving as references. In Stage II, human- and LLM-based ratings were entered into the same constrained optimization procedure. In Stage III, 400 participants completed the resulting forms, and latent trait estimates were compared using graded Thurstonian item response theory. LLM ratings showed strong item-level correspondence with human ratings (sentence: $r = 0.942$–0.963; adjective: $r = 0.979$–0.980), but optimized forms showed partial item overlap and sparse exact pair overlap. Despite these design differences, LLM-based forms produced latent trait estimates that closely matched those from human-based forms (mean $r = 0.815$ for sentence items and 0.768 for adjective items). These results support the conditional use of LLM-generated ratings as scalable construction inputs while showing why downstream validation is necessary.[1]

1 Introduction

Large language models (LLMs) are increasingly used to produce ratings and other evaluative judgments that would otherwise be collected from human annotators. Research on LLM-based annotation and LLM-as-a-judge evaluation has examined whether model-produced judgments align with human judgments, while also documenting biases and instability in such outputs (Gu et al., 2026; Liu et al., 2023; Zheng et al., 2023). Within this literature, human–model agreement often serves as a starting point for broader claims that model-generated judgments can reduce reliance on human annotation in specified tasks (Gilardi et al., 2023). Such claims require evidence beyond judgment-level agreement when model-generated judgments shape downstream outputs such as datasets, evaluation scores, selected examples, or analysis results. This motivates a broader methodological question: in a specified downstream pipeline, does replacing human judgment inputs with model-generated ones change the resulting outputs or inferences? If such replacement is warranted, it could make rating-dependent research workflows easier to scale, repeat, and adapt while reducing the cost and delay of collecting human ratings.

We examine this problem in forced-choice psychometric questionnaire construction, a focused downstream setting where item-level social-desirability ratings are a key construction input. Psychometric questionnaires are central tools for studying personality traits, values, and attitudes, and they are increasingly used in NLP to probe language-model behavior (Huang et al., 2024). Forced-choice questionnaires are widely used to reduce faking and socially desirable responding in evaluative settings (Cao and Drasgow, 2019; Speer et al., 2023); applied instruments include the Tailored Adaptive Personality Assessment System (TAPAS) for U.S. military accession and classification (Stark et al., 2014) and OPQ-family instruments for personnel selection (Brown and Bartram, 2009). Recent NLP work has also used forced-choice questionnaires to evaluate LLMs as respondents and mitigate socially desirable responding (Okada et al., 2026). In a forced-choice questionnaire, respondents are presented with a response

[1]Data and code available at: https://osf.io/gmdf4/overview

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 3991–4020
November 6-10, 2026 ©2026 Association for Computational Linguistics

block containing two or more items, and the items in that block serve as the alternatives to be compared. In the two-item blocks used in the present study, respondents indicate which item in the pair describes them more accurately and by how much, rather than rating each item independently, as illustrated in Figure 1(c). To reduce desirability-driven choices, candidate items are paired across domains to have similar levels of social desirability, defined as the perceived favorability of endorsement. Thus, constructing desirability-matched pairs requires item-level social-desirability ratings. Here, item social desirability is distinct from socially desirable responding, which refers to a response tendency among respondents. Because collecting item-level social-desirability ratings from human raters is resource-intensive, replacing human rating collection with LLM-generated ratings could reduce the resources required for scale construction. We therefore examine whether LLMs can substitute for human raters in providing item-level social-desirability ratings for this purpose. Accordingly, as an instance of downstream validation, the present study examines not only rating-level correspondence between LLM-generated and human ratings, but also how well this correspondence is preserved in downstream outcomes, including the constructed forced-choice forms and human measurement outcomes. LLMs were used only as item-level social-desirability raters. No LLM answered a questionnaire or received a personality-trait estimate; all Stage III respondents were human participants.

This setting creates substantial *researcher degrees of freedom* (Simmons et al., 2011), spanning the choice of LLM, prompt wording, output aggregation, item exclusions, optimization constraints, scoring model, and criteria for comparing latent trait estimates. We therefore treat the substitutability of LLM-generated rating inputs as a *preregistered* downstream validation problem rather than a post hoc agreement analysis. Following calls to make preregistration more explicit in NLP studies (van Miltenburg et al., 2021), our preregistration[2] specifies the rating procedure, aggregation rules, optimization pipeline, human-subject validation, exclusion criteria, scoring model, and confirmatory decision criteria before the relevant outcomes are observed (Nosek et al., 2018). We instantiate this framework using the HEXACO (Honesty-Humility,

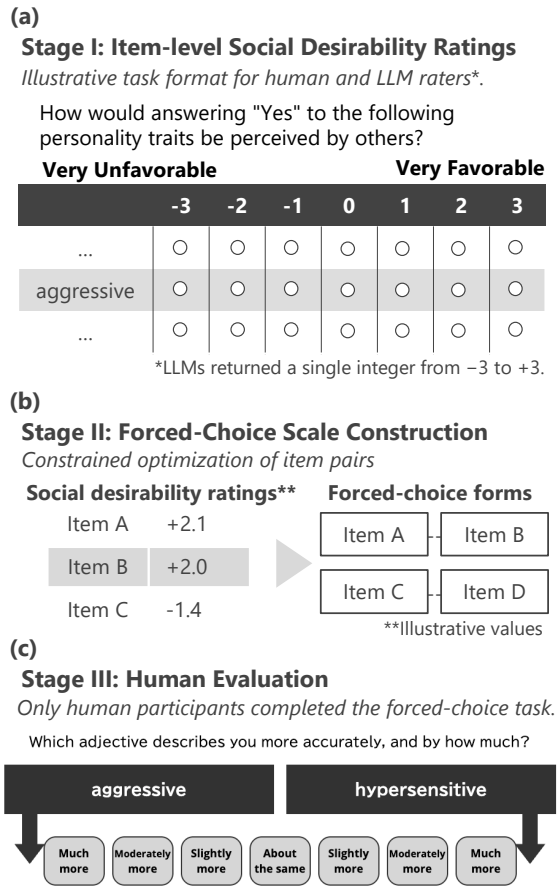


Figure 1: From item-level social-desirability ratings to human forced-choice responses. (a) In Stage I, human and LLM raters evaluated items individually for social desirability on a -3 to +3 perceived-favorability scale. (b) In Stage II, these ratings were used to construct forced-choice forms by constrained optimization. (c) In Stage III, human participants indicated which item in each pair described them more accurately and by how much using a seven-category graded forced-choice scale. Adjective-format examples are shown; the sentence-format condition followed the same procedures.

Emotionality, Extraversion, Agreeableness, Conscientiousness, Openness) model of personality, a six-dimensional lexical model closely related to the Big Five and accompanied by publicly available measurement resources (Ashton and Lee, 2005; Ashton et al., 2007).

Figure 2 summarizes the preregistered three-stage validation design. To test whether LLM-generated rating inputs remain useful beyond item-level agreement, we compare the human-rating-based and LLM-rating-based pipelines at three levels: input-level rating correspondence (Stage I), design-level agreement in optimization-based forced-choice scale construction (Stage II), and

[2]Preregistration: https://osf.io/xca6z/overview

measurement-level correspondence of latent trait estimates from human respondents (Stage III).

Empirically, we find that LLM-generated social-desirability ratings closely track human ratings at the item level, and the resulting forced-choice forms yield strongly corresponding latent trait estimates despite different item-pair configurations. These findings suggest that LLM-generated ratings can serve as inputs to forced-choice questionnaire construction, while showing why rating-level agreement alone is insufficient for evaluating replacement claims.

This paper contributes (i) a downstream validation framing for LLM-generated psychological ratings, (ii) a preregistered pipeline separating input-level rating correspondence, design-level agreement in scale construction, and measurement-level correspondence, and (iii) a randomized human-subject test of whether LLM-rating-based forced-choice forms yield latent trait estimates comparable to those from human-rating-based forms under graded Thurstonian IRT.

2 Related Work

LLM-generated judgments and downstream validation. LLMs are increasingly used as sources of labels, scores, rankings, and evaluative judgments in NLP workflows. Work on LLM-based annotation and LLM-as-a-judge evaluation has examined alignment with human judgments, biases, instability, and task dependence in model outputs (Gu et al., 2026; Liu et al., 2023; Zheng et al., 2023). However, agreement with human judgments does not by itself establish that LLM-generated judgments can be used without affecting downstream results. Recent work has therefore examined how LLM-generated or LLM-assisted annotations affect subsequent statistical inference and model evaluation. For example, Gligorić et al. (2025) developed a framework for combining LLM and human annotations while preserving the validity of downstream statistical conclusions, whereas Schroeder et al. (2025) showed that LLM-assisted annotation can alter label distributions and consequently affect conclusions about model performance. These studies motivate evaluating LLM-generated judgments not only by their correspondence with human judgments, but also by the transformations and endpoints involved in their intended use. Whereas prior work has examined downstream statistical inference or model evaluation after LLM-assisted annotation, the present study examines constrained questionnaire construction as the transformation and human latent trait estimates as the endpoint.

Social-desirability ratings and forced-choice scale construction. The judgments most relevant here are psychological ratings of item content. Personality research has long studied normative properties of trait descriptors, including social desirability and valence, because such properties shape how items are perceived and answered. Recent studies have provided large-scale ratings of personality descriptors and examined whether machine-based methods, including LLMs, approximate human ratings of socially relevant attributes (Britz et al., 2023; Hommel, 2023; Leistner et al., 2025). Existing work in this area, however, has mainly evaluated rating-level correspondence. Forced-choice questionnaires ask respondents to compare alternatives within blocks and are used to reduce faking, acquiescence, and socially desirable responding in evaluative assessment settings (Cao and Drasgow, 2019; Martínez and Salgado, 2021; Speer et al., 2023). Their construction requires desirability matching: items within a block should be similar in perceived favorability while measuring different domains (Brown, 2016; Pavlov et al., 2021). Comparative responses are then modeled with item response theory, including Thurstonian approaches, to recover normative latent trait estimates (Brown and Maydeu-Olivares, 2011; Brown, 2016; Bürkner, 2022). Recent NLP work has used forced-choice questionnaires to evaluate LLMs as respondents (Okada et al., 2026). We examine a different role for LLMs: providing item-level social-desirability ratings used to construct forced-choice questionnaires for human respondents.

Positioning and preregistration of the present study. Evaluating LLM-generated ratings as research inputs requires more than high correlation with human ratings. In measurement research, convergent validity between two measures does not ensure equivalent downstream results (Carlson and Herdman, 2012). In the present setting, LLM-generated social-desirability ratings are transformed through a constrained questionnaire-construction procedure. We therefore evaluate not only rating-level correspondence, but also the resulting item and pair selections and the latent trait estimates obtained when the constructed forms are administered to human participants. This validation problem involves substantial researcher de-

grees of freedom, including model choice, prompt wording, aggregation, exclusions, optimization constraints, scoring models, and comparison criteria. Preregistration helps distinguish confirmatory tests from analyses chosen after observing outcomes and has been recommended for NLP studies involving human evaluation and flexible analysis pipelines (Simmons et al., 2011; Nosek et al., 2018; van Miltenburg et al., 2021). We therefore treat LLM-rating substitutability as a preregistered downstream validation problem, testing whether correspondence at the rating level extends through forced-choice scale construction and, ultimately, to human latent trait estimates.

3 Method

3.1 Overview

We implemented the preregistered three-stage pipeline shown in Figure 2. Sentence- and adjective-format item pools were analyzed separately throughout. In all primary comparisons, outputs derived from LLM-generated desirability ratings were compared with outputs derived from human desirability ratings under matched item pools, optimization constraints, administration procedures, and scoring models. The human-rating-based pipeline served as the reference baseline, and this baseline was compared with otherwise identical pipelines that used desirability ratings generated by GPT, Gemini, or Claude.

3.2 Item Pools

We used two English HEXACO-based item pools as the starting materials for the study: 60 sentence-format items from the HEXACO-60 inventory (Ashton and Lee, 2009) and 60 adjective-format items from the HEXACO Adjective Scales (Romano and Perugini, 2023; Romano et al., 2023). Both item pools covered the six HEXACO domains—Honesty–Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness to Experience—and were used throughout the three-stage pipeline. Each item was associated with a domain label and keying direction, which were used in the Stage II forced-choice scale-construction procedure.

3.3 Stage I: Item-level Social Desirability Ratings

Stage I compared LLM-generated social desirability ratings with fixed human-reference ratings for the same item pools. The human-reference ratings were obtained from a pre-existing human-rating dataset collected by the authors in July 2025, before the preregistration of the present study, using an online survey with participants recruited through Prolific; additional details are provided in Appendix A. The preregistered protocol for the present study treated this dataset as fixed and specified its use as the human-reference benchmark for Stage I and as the human-rating input for Stage II scale construction.

Participants in the prior rating study were required to be native English speakers aged between 18 and 69 years. They were randomly assigned to either the sentence-format condition

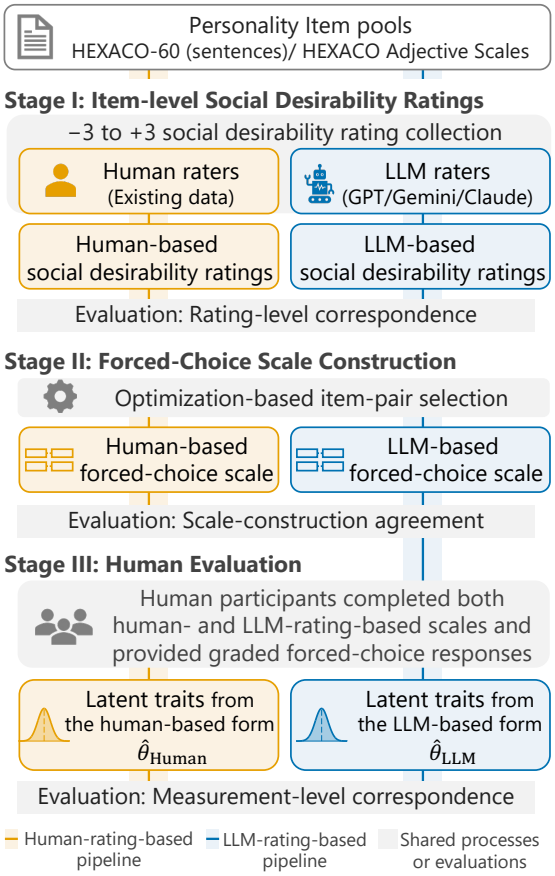


Figure 2: Overview of the three-stage evaluation framework used in the present study. Stage I evaluated rating-level correspondence between LLM-based and human-based social desirability ratings. Stage II evaluated agreement between forced-choice scales constructed from the two rating sources using the same optimization procedure. Stage III evaluated measurement-level correspondence between latent trait estimates derived from the resulting scales. Sentence-format and adjective-format item pools were analyzed separately throughout all stages.

or the adjective-format condition and rated all 60 items in their assigned format. For each item, participants rated how answering "Yes" to the personality item would be perceived by others, using a 7-point scale from -3 (very unfavorable) to $+3$ (very favorable). The final human-rating dataset included responses from 277 participants for the sentence format and 321 participants for the adjective format. Item-level mean human ratings were computed and used as the fixed human-reference desirability estimates. Additional details of the prior human-rating data collection are provided in Appendix A.

LLM-based desirability ratings were collected for the same items using the same conceptual rating task and response scale. We evaluated three high-capacity API models: GPT-5.4, Gemini 2.5 Pro, and Claude Opus 4.6. Each API request contained a single item and used a fixed prompt that mirrored the human-rating instruction and used the same seven-point -3 to $+3$ response scale. Following the preregistered protocol, each item was evaluated 15 times per model and scale format. Returned responses were treated as valid only if they consisted of a single integer within the allowed response range, and valid responses were averaged to obtain one LLM-based desirability estimate per item, model, and scale format. The default generation settings of each provider were used; no persona assignment, prompt optimization, or hyperparameter tuning was performed. Full prompt text, response-processing rules, model identifiers, API settings, and model snapshots are provided in Appendix B.

Stage I analyses compared item-level mean LLM ratings with item-level mean human-reference ratings for each model and scale format.

3.4 Stage II: Forced-Choice Scale Construction

In Stage II, we constructed graded forced-choice item-pair inventories separately for each scale format and rating source. Optimization-based assembly has been used in prior forced-choice questionnaire construction, including genetic algorithms (Kreitchmann et al., 2022), simulated annealing as implemented in the `autoFC` package (Li et al., 2022), and mixed-integer optimization (Frick, 2023). Such inventories require item pairs to be closely matched on social desirability while simultaneously satisfying whole-form structural constraints such as domain balance and keying balance. Accordingly, Stage II served as both a scale-construction step and a design-level comparison of the resulting item sets and pair configurations; measurement-level correspondence was evaluated downstream in Stage III.

Following Okada et al. (2026), pair selection was formulated as a constrained mixed-integer lexicographic optimization problem. For each source and scale format, the optimization selected exactly 21 cross-domain pairs, matching items on social desirability while preserving structural balance. It first minimized the maximum within-pair desirability mismatch and then, subject to that minimax optimum, minimized the total squared mismatch across selected pairs. Structural constraints enforced item uniqueness, exactly seven selected items per HEXACO domain, one or two pairs per unordered domain combination, 40–60% mixed-key pairs, and at least 30% representation of each keying direction within each domain. Deterministic tie-breaking yielded a unique solution. This procedure produced one human-based forced-choice inventory and one LLM-based inventory for each model within each scale format. Full mathematical details, solver settings, and tie-breaking procedures are provided in Appendix C.

3.5 Stage III: Human Evaluation

In Stage III, we administered the constructed forced-choice inventories to human participants recruited through Prolific. The final analytic sample comprised 400 participants with approved Prolific submissions. Eligible participants were required to be native English speakers aged 18–69 years and to complete the study on a desktop or laptop computer. Participants were randomly assigned to the sentence-format condition ($N = 214$) or adjective-format condition ($N = 186$).

Within each scale-format condition, participants completed both the human-based inventory and the LLM-based inventory carried forward to Stage III. The presentation order of the two inventories was counterbalanced across participants, and item-pair order was randomized within each inventory. Participants responded to each item pair on a 7-category bipolar scale indicating which option described them more accurately and by how much.

Responses were scored with a graded Thurstonian item response model (graded TIRT model; Brown, 2016; Brown and Maydeu-Olivares, 2011; Bürkner, 2022; Okada and Bunji, 2021). The model treats each forced-choice response as an

ordinal observation of the latent utility difference between the two items in a pair. For participant n and pair p, the response $Y_{np} \in \{1, \ldots, 7\}$ was modeled as

$$Y_{np} \sim \text{OrderedLogistic}(\eta_{np}, \mathbf{c}_p),$$

where $\mathbf{c}_p$ denotes the ordered thresholds for the seven response categories. The latent utility difference was defined as

$$\eta_{np} = \frac{1}{\sqrt{2}} \left(a_{R(p)} z_{nR(p)} - a_{L(p)} z_{nL(p)} \right),$$

where $L(p)$ and $R(p)$ index the left and right items in pair p, a_j is the signed item discrimination parameter, and z_{nj} is the latent utility of item j for participant n. Each item loaded on a single HEXACO domain, such that $z_{nj} = \boldsymbol{\theta}_n^\top \mathbf{q}_j$, where $\boldsymbol{\theta}_n$ is the participant's six-dimensional latent trait vector and $\mathbf{q}_j$ is the item-domain loading vector. Item keying was incorporated by parameterizing discrimination as $a_j = g_j a_j^+$, where $g_j \in \{-1, +1\}$ denotes the keying direction and $a_j^+ > 0$ denotes the discrimination magnitude.

For each item format, the human-based and LLM-based inventories were combined into a joint pair bank and calibrated together. Because each inventory contained 21 forced-choice pairs, the joint pair bank contained up to 42 pair instances per item format. To avoid imposing equality constraints across forms, items originating from different scale sources were treated as distinct item instances during calibration, even when their wording overlapped across inventories. After joint calibration, form-specific scoring was conducted separately for the human-based and LLM-based inventories while fixing the calibrated item parameters, thresholds, and latent-trait correlation structure. This procedure yielded separate human-based and LLM-based latent trait estimates for each participant on a common metric, which were then compared in the Stage III confirmatory analyses. TIRT estimation, including the joint calibration and subsequent form-specific scoring steps, was implemented in Stan using Hamiltonian Monte Carlo sampling via `cmdstanr` (Carpenter et al., 2017; Gabry et al., 2024). Priors, sampler settings, and additional model specifications are provided in Appendix D.

Preregistered exclusion criteria included failed attention checks, straightlining, implausibly fast responses, and incomplete responses; no approved participant met the exclusion criteria. Recruitment details, instructions, screening procedures, and sample characteristics are reported in Appendix E.

3.6 Analysis Overview

Primary analyses followed the preregistered plan. In Stage I, we computed Pearson correlations between item-level mean human ratings and item-level mean LLM ratings for each model and scale format. In Stage II, we quantified scale-construction agreement using item-level and pair-level agreement rates and Jaccard coefficients between the human-based and LLM-based selected sets. In Stage III, we computed Pearson correlations between participant-level latent trait estimates from the human-based and LLM-based forms for each HEXACO domain and summarized them by Fisher z-averaging across domains. The preregistered criteria treated $r \geq .50$ as the minimum level of hypothesis support and $r \geq .70$ as sufficient evidence of correspondence for the Stage I and Stage III correlation-based outcomes. These thresholds were adopted as convergent-validity benchmarks based on Carlson and Herdman (2012), who recommended convergent validity coefficients above $r = .70$ and advised against coefficients below $r = .50$. Stage II overlap analyses were used to characterize design-level propagation rather than as a strict equivalence criterion. Exploratory analyses and the mapping between preregistered and reported analyses are provided in Appendix F.

4 Results

We report the results of the preregistered confirmatory analyses. All primary outcomes correspond to the indices specified in the preregistration. Exploratory analyses are reported separately and are not used to evaluate the preregistered hypotheses.

4.1 Stage I: Rating-Level Correlation

To assess rating-level correspondence, we computed Pearson correlation coefficients between item-level mean human-reference and LLM ratings for each scale format and LLM. For adjective items, Item 44 was excluded because of a wording discrepancy between the HEXACO website version used to collect both rating sets (*sicure*) and the published HAS appendix (*secure*; see Appendix F.6). Adjective-format analyses therefore used 59 rather than the preregistered 60 items.

As shown in Table 1, correlations between LLM-generated and human-reference ratings were uni-

Model	r	95% CI	N
Sentence			
GPT	0.942***	[0.904, 0.965]	60
Gemini	0.948***	[0.915, 0.969]	60
Claude	0.963***	[0.939, 0.978]	60
Adjective			
GPT	0.980***	[0.967, 0.988]	59
Gemini	0.980***	[0.967, 0.988]	59
Claude	0.979***	[0.964, 0.987]	59

Table 1: Correlations between LLM and human ratings. r = Pearson correlation; N = number of items. Models: GPT-5.4, Gemini 2.5 Pro, and Claude Opus 4.6. $^{**}p < 0.05$, $^{***}p < 0.001$.

formly high across models and scale formats. For sentence items, correlations ranged from 0.942 to 0.963, whereas for adjective items, correlations ranged from 0.979 to 0.980. All correlations exceeded the preregistered Stage I support criterion ($r \geq 0.70$), supporting the hypothesis that LLM-generated social-desirability ratings show strong item-level correspondence with human ratings. Pearson correlation was the preregistered primary Stage I index because the Stage I question concerned linear correspondence in item locations on the perceived-favorability dimension; absolute agreement and response-scale use were assessed separately using descriptive statistics and concordance correlation coefficients (CCCs). These complementary analyses showed that item-level mean ratings from the LLMs spanned a wider range of the response scale than those from human raters: the observed ranges were 6.000 for all LLMs, compared with 4.148 for human sentence ratings and 4.969 for human adjective ratings, and the corresponding CCCs were .804–.871 and .915–.928 (Appendix G). Following the preregistered rule of selecting the LLM with the highest Stage I correlation with human ratings within each format, we carried forward Claude for sentences and GPT for adjectives to Stage III. For adjectives, GPT and Gemini both rounded to $r = .980$, but GPT had the higher unrounded correlation.

4.2 Stage II: Pair Construction Agreement

Table 2 contrasts overlap in selected item sets with exact item-pair reconstruction. Because Stage II was designed to characterize design-level propagation rather than strict equivalence, we interpret these indices descriptively. At the item level, agreement rates ranged from 0.667 to 0.786, and Jaccard coefficients ranged from 0.500 to 0.647. These agreement rates correspond to 28–33 shared selected items out of the 42 items in each inventory. For sentence items, Gemini and Claude tied for the highest item-level agreement (agreement = 0.786, Jaccard = 0.647); for adjective items, Gemini showed the highest item-level agreement (agreement = 0.738, Jaccard = 0.585), followed by GPT (agreement = 0.714, Jaccard = 0.556), despite GPT's higher Stage I correlation.

At the pair level, agreement was much lower. Pair agreement rates ranged from 0.000 to 0.238, and Jaccard coefficients ranged from 0.000 to 0.135, corresponding to 0–5 shared exact item pairs out of 21. For sentence items, Claude showed the highest pair-level agreement (agreement = 0.238, Jaccard = 0.135). For adjective items, GPT was the only model with non-zero pair-level agreement (agreement = 0.048, Jaccard = 0.024).

Thus, LLM-based and human-based ratings produced partially overlapping selected item sets but substantially different exact pair configurations. The models selected in Stage I showed the strongest pair-level agreement within their respective item formats, yet exact pair overlap remained sparse. These results indicate that high rating-level correspondence can partly propagate to item selection without guaranteeing identical optimized pair configurations, motivating the Stage III test of latent-trait correspondence.

As an exploratory Human–Human benchmark, exact-pair overlap was also low (mean Jaccard = .027 for sentence items and .045 for adjective items), and none of the six observed Human–LLM Jaccard coefficients fell below the corresponding Human–Human 2.5th percentile. Thus, none of the observed Human–LLM pair-overlap values was unusually low relative to the corresponding resampled Human–Human distribution; this comparison was exploratory and descriptive rather than a formal equivalence test (Appendix H).

Additional exploratory analyses examined optimization sensitivity and solution multiplicity to further interpret the low exact-pair overlap observed in Stage II (Appendix H). Appendix I also reports analogous Stage I and Stage II analyses for two open-weight models, Llama 3.3 70B Instruct and Mistral Large 3, using the same rating-comparison and scale-construction framework. Because these open-weight analyses did not include Stage III human-participant validation, they address only rat-

	Item		Pair	
Model	Agr.	Jac.	Agr.	Jac.
Sentence				
GPT	0.690	0.527	0.048	0.024
Gemini	0.786	0.647	0.095	0.050
Claude	0.786	0.647	0.238	0.135
Adjective				
GPT	0.714	0.556	0.048	0.024
Gemini	0.738	0.585	0.000	0.000
Claude	0.667	0.500	0.000	0.000

Table 2: Item- and pair-level agreement between LLM-based and human-based scales. Agr. = agreement rate (proportion of elements in the human-based scale that were also included in the corresponding LLM-based scale); Jac. = Jaccard coefficient. Item-level indices were computed based on the set of selected items, and pair-level indices were computed based on the set of selected item pairs.

ing correspondence and propagation through scale construction.

4.3 Stage III: Measurement-Level Correspondence

For the Claude-based sentence inventory and GPT-based adjective inventory carried forward to Stage III, we evaluated measurement-level correspondence with the corresponding human-based inventories. The TIRT models showed satisfactory convergence, with a maximum $\hat{R}$ of 1.023 and no divergent transitions. We computed Pearson correlations between participant-level latent trait estimates from the human-based and LLM-based forms for each HEXACO domain and summarized the domain-specific correlations by Fisher z-averaging.

As shown in Table 3, correlations were positive across all HEXACO domains for both scale formats. For sentence items, domain-specific correlations ranged from 0.781 to 0.844, with a mean r of 0.815. For adjective items, correlations ranged from 0.659 to 0.828, with a mean r of 0.768. The preregistered format-level mean correlations exceeded the $r \geq 0.70$ criterion in both item formats, although the adjective-format Honesty–Humility correlation fell slightly below 0.70 ($r = 0.659$). Figure 3 visualizes these participant-level associations. In both formats, latent trait estimates from the selected LLM-based scales showed a clear positive linear association with those from the human-based inventories.

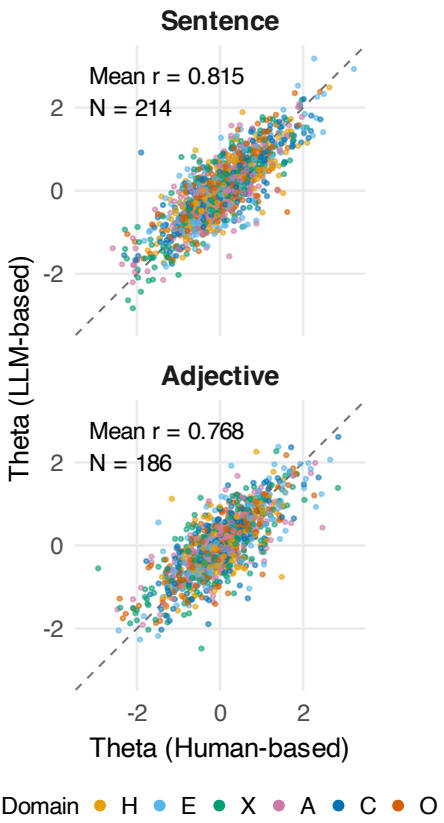


Figure 3: Scatter plots of latent trait estimates from the human-based and LLM-based scales for the sentence (top) and adjective (bottom) formats. Each point represents one participant-by-domain latent trait estimate; each participant therefore contributes six points. Colors indicate the six HEXACO domains. The reported mean r is computed by Fisher z averaging the six domain-specific correlations rather than by pooling all plotted points. N denotes the number of participants in each format condition. The dashed line indicates equality.

In addition to the overall correspondence observed at the participant level, Figure 4 illustrates the domain-level latent trait profiles derived from the two types of scales. Across both scale formats, the profiles obtained from the LLM-based scales closely matched those from the human-based scales, showing highly similar patterns across HEXACO domains. This suggests that the agreement between the two approaches extends beyond overall association to the structural pattern of trait estimates across domains.

When comparing scale formats, sentence items tended to show slightly higher correlations across most HEXACO domains, although adjective items showed a higher correlation for Openness. This pattern is consistent with the Stage II scale-

	Domain-specific correlations						Summary	
Scale format	H	E	X	A	C	O	Mean	N
Sentence	0.781***	0.830***	0.844***	0.783***	0.839***	0.806***	0.815	214
Adjective	0.659***	0.825***	0.761***	0.737***	0.762***	0.828***	0.768	186

Table 3: Correlations between latent trait estimates derived from human-based and LLM-based scales. Values are Pearson correlations between human- and LLM-based latent trait estimates; stars indicate two-sided tests of domain-specific correlations only (**$p < .05$, ***$p < .001$). Mean r is the back-transformed Fisher-z average across domains and is not assigned a star; N is the number of participants.

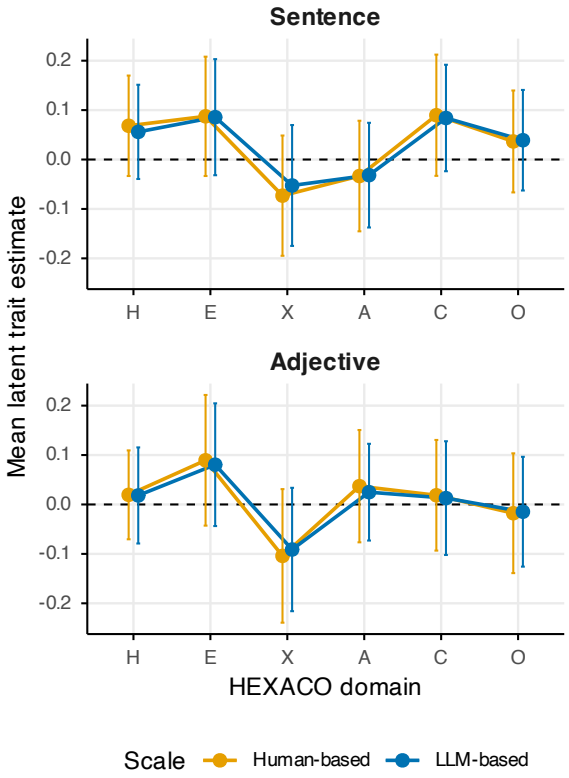


Figure 4: Mean latent trait profiles across HEXACO domains for the human-based and LLM-based scales. Points and connecting lines represent mean latent trait estimates for each domain, with error bars indicating 95% confidence intervals.

construction results (Table 2), where the sentence-format model carried forward to Stage III (Claude) showed higher item- and pair-level agreement with the human-based scale than the corresponding adjective-format model (GPT).

Overall, scales constructed from LLM ratings yielded latent trait estimates that strongly corresponded to those from human-based scales. Additional exploratory diagnostics of Stage III agreement, including CCCs, mean differences, posterior SD summaries, and shared-item parameter stability, are reported in Appendix J.

5 Conclusion

This study evaluated whether LLM-generated social-desirability ratings can serve as construction inputs for forced-choice personality scales. We used a preregistered three-stage validation pipeline linking rating-level correspondence, optimization-based pair selection, and measurement-level correspondence in latent trait estimates from human respondents.

The results show that LLM-generated desirability ratings closely tracked human-reference ratings across models and scale formats. At the construction stage, however, high item-level correlations did not yield exact reconstruction of the human-based forced-choice forms: item overlap was partial, and pair-level overlap remained limited. Even with these differences in item-pair configuration, the selected LLM-based forms produced latent trait estimates that strongly corresponded to those from the human-based forms. Thus, downstream measurement correspondence was achievable despite differences in the specific item pairs selected by the construction algorithm.

These findings support the conditional use of LLM-generated ratings as scalable inputs for forced-choice scale construction: rating-level agreement is useful, but replacement claims require evidence that model-generated judgments preserve downstream artifacts and measurement outcomes. More broadly, the framework offers a template for evaluating LLM-generated judgments in research pipelines that use model outputs to build datasets, select examples, compute scores, or guide later analyses. Under controlled conditions, LLM ratings may reduce the burden of rating collection, provided that the downstream pipeline has been validated for the intended measurement purpose. Future work could use these questionnaires to examine socially desirable responding in LLMs as respondents.

6 Limitations

Several limitations of the present study should be noted. First, the present findings are bounded by the specific setting examined here: forced-choice questionnaire construction using English HEXACO-based item pools, the present social-desirability rating task, and the specified optimization framework. Accordingly, the results may not generalize to other languages, psychological constructs, item formats, or scale-construction tasks. Future studies should examine whether similar patterns of correspondence are observed in other psychometric contexts.

Second, the results may depend on the specific LLMs, model versions, and prompting procedures used in this study. Because LLM behavior can vary across systems and may change over time as models are updated, the present findings should not be interpreted as fixed characteristics of any particular model family.

Third, although human ratings were treated as the primary benchmark in the present study, human judgments themselves are not free from variability or bias. Desirability ratings may vary across raters, cultural contexts, and assessment conditions. Accordingly, correspondence with human ratings should not be interpreted as correspondence with an objective ground truth.

Fourth, although substantial correspondence was observed at the latent trait level, the present results do not establish strict equivalence or interchangeability between human-based and LLM-based scales. In particular, overlap at the level of item-pair selection was relatively limited, and individual-level discrepancies remained observable in exploratory analyses. Additional validation studies are therefore needed before LLM-based approaches can be considered interchangeable with conventional human-based procedures. Without sufficient validation or human oversight, misuse or overreliance on LLM-assisted psychometric methods could lead to inappropriate interpretations or decisions in applied settings.

Finally, the present study focused primarily on correspondence between human- and LLM-based outcomes and did not directly evaluate broader aspects of psychometric quality. For example, the present analyses did not examine whether LLM-based scales predict external criteria or real-world behavior to the same extent as human-constructed scales, whether they function equivalently across demographic or cultural groups, or whether their measurement properties remain stable over time. These issues remain important directions for future research.

7 Ethical Considerations

This study used human participant data collected through questionnaire-based personality assessment procedures. All participants provided informed consent prior to participation, and responses were collected anonymously through an online survey platform. The pre-existing human-rating dataset was reviewed and approved by the Human Ethics Committee of the Graduate School of Business Administration, Kobe University (Approval No. 2025-2). The Stage III forced-choice survey was reviewed and approved by the Research Ethics Committee of the University of Tokyo (Approval No. E2025ALS460). The data included anonymized desirability ratings and forced-choice questionnaire responses, and no personally identifying information was intentionally collected as part of the study procedures. The study materials were limited to personality-related questionnaire items and were not expected to contain offensive or harmful content beyond ordinary everyday language.

The present study examined the potential utility of LLM-generated desirability ratings for psychometric scale construction under controlled research conditions. Although substantial correspondence was observed between human- and LLM-based outcomes, the findings should not be interpreted as supporting fully automated psychological assessment or the replacement of human expertise in psychometric practice.

In particular, the present findings are not intended to justify the use of LLM-based personality assessment in high-stakes contexts such as personnel selection, educational evaluation, or clinical decision making. The present study was designed as a methodological investigation of LLM-assisted scale construction, and additional validation research remains necessary before such approaches can be considered for practical deployment in applied settings. Imperfect agreement in item-pair selection and individual-level latent trait estimates underscores the need for human oversight in LLM-assisted psychometric applications.

Acknowledgments

AI-based assistants (including large language model systems) were used during portions of the coding and manuscript-preparation process, including code drafting, language editing, and organizational support. All generated content was reviewed and revised by the authors, who take full responsibility for the final manuscript, analyses, and interpretations. This work was supported by JSPS KAKENHI Grant Numbers 25H00577, 25K06746 and JST A-STEP Grant Number JPMJTR26R3.

References

Michael C. Ashton and Kibeom Lee. 2005. Honesty-Humility, the Big Five, and the Five-Factor Model. *Journal of Personality*, 73(5):1321–1354.

Michael C. Ashton and Kibeom Lee. 2009. The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment*, 91:340–345.

Michael C. Ashton, Kibeom Lee, and Lewis R. Goldberg. 2007. The IPIP–HEXACO scales: An alternative, public-domain measure of the personality constructs in the HEXACO model. *Personality and Individual Differences*, 42(8):1515–1526.

Sara Britz, Lena Rader, Siegfried Gauggel, and Verena Mainz. 2023. An English list of trait words including valence, social desirability, and observability ratings. *Behavior Research Methods*, 55:2669–2686.

Anna Brown. 2016. Item response models for forced-choice questionnaires: A common framework. *Psychometrika*, 81:135–160.

Anna Brown and Dave Bartram. 2009. The occupational personality questionnaire revolution: Applying item response theory to questionnaire design and scoring. Technical report, SHL Group, Thames Ditton, Surrey.

Anna Brown and Alberto Maydeu-Olivares. 2011. Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, 71:460–502.

Paul-Christian Bürkner. 2022. On the information obtainable from comparative judgments. *Psychometrika*, 87:1439–1472.

Mengyang Cao and Fritz Drasgow. 2019. Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *Journal of Applied Psychology*, 104(11):1347–1368.

Kevin D. Carlson and Andrew O. Herdman. 2012. Understanding the impact of convergent validity on research results. *Organizational Research Methods*, 15:17–32.

Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–32.

Susanne Frick. 2023. Estimating and using block information in the Thurstonian IRT model. *Psychometrika*, 88(4):1556–1589.

Jonah Gabry, Rok Češnovar, Andrew Johnson, and Steve Bronder. 2024. *cmdstanr: R Interface to CmdStan*. R package version 0.8.0.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Kristina Gligorić, Tijana Zrnic, Cinoo Lee, Emmanuel Candès, and Dan Jurafsky. 2025. Can unconfident LLM annotations be used for confident conclusions? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3514–3533, Albuquerque, New Mexico. Association for Computational Linguistics.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Zhouchi Lin, Bowen Zhang, Lionel Ni, Wen Gao, Yuanzhuo Wang, and Jian Guo. 2026. A survey on LLM-as-a-Judge. *The Innovation*, 7(6):101253.

Gurobi Optimization, LLC. 2026. Gurobi Optimizer Reference Manual. Version 13.0-1.

Björn E. Hommel. 2023. Expanding the methodological toolbox: Machine-based item desirability ratings as an alternative to human-based ratings. *Personality and Individual Differences*, 213:112307.

Jen-tse Huang, Wenxiang Jiao, Man Ho Lam, Eric John Li, Wenxuan Wang, and Michael Lyu. 2024. On the reliability of psychological scales on large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6152–6173, Miami, Florida, USA. Association for Computational Linguistics.

Rodrigo Schames Kreitchmann, Francisco J. Abad, and Miguel A. Sorrel. 2022. A genetic algorithm for optimal assembly of pairwise forced-choice questionnaires. *Behavior Research Methods*, 54(3):1476–1492.

Moritz Leistner, Björn E. Hommel, Leon P. Wendt, and Daniel Leising. 2025. Properties of person descriptors in the natural German language: A preregistered replication and extension. *Personality Science*, 6:1–13.

Mengtong Li, Tianjun Sun, and Bo Zhang. 2022. autoFC: An R package for automatic item pairing in forced-choice test construction. *Applied Psychological Measurement*, 46(1):70–72.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Alexandra Martínez and Jesús F. Salgado. 2021. A meta-analysis of the faking resistance of forced-choice personality inventories. *Frontiers in Psychology*, 12:732241.

Joris Mulder, Donald R. Williams, Xin Gu, Andrew Tomarken, Florian Böing-Messing, Anton Olsson-Collentine, Marlyne Meijerink-Bosman, Janosch Menke, Robbie van Aert, Jean-Paul Fox, Herbert Hoijtink, Yves Rosseel, Eric-Jan Wagenmakers, and Caspar van Lissa. 2021. BFpack: Flexible Bayes factor testing of scientific theories in R. *Journal of Statistical Software*, 100(18):1–63.

Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. 2018. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606.

Kensuke Okada and Kyosuke Bunji. 2021. Increase of reliability by incorporating response time into the paired-comparison psychological measurement. *Behaviormetrika*, 48:169–177.

Kensuke Okada, Yui Furukawa, and Kyosuke Bunji. 2026. Quantifying and mitigating socially desirable responding in LLMs: A desirability-matched graded forced-choice psychometric study. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 40148–40166, San Diego, California, United States. Association for Computational Linguistics.

Goran Pavlov, Dexin Shi, Alberto Maydeu-Olivares, and Amanda Fairchild. 2021. Item desirability matching in forced-choice test construction. *Personality and Individual Differences*, 183:111114.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Daniele Romano, Giulio Costantini, Juliette Richetin, and Marco Perugini. 2023. The HEXACO adjective scales and its psychometric properties. *Assessment*, 30(8):2510–2532.

Daniele Romano and Marco Perugini. 2023. Initial validation of the English HEXACO adjective scales (HAS). HEXACO-PI-R website. `https://hexaco.org/downloads/Eng_HAS_InitialValidation(OSF)_MANUSCRIPT.pdf`.

Hope Schroeder, Deb Roy, and Jad Kabbara. 2025. Just put a human in the loop? Investigating LLM-assisted annotation for subjective tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25771–25795, Vienna, Austria. Association for Computational Linguistics.

Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366.

Andrew B. Speer, Lauren J. Wegmeyer, Andrew P. Tenbrink, Angie Y. Delacruz, Neil D. Christiansen, and Rouan M. Salim. 2023. Comparing forced-choice and single-stimulus personality scores on a level playing field: A meta-analysis of psychometric properties and susceptibility to faking. *Journal of Applied Psychology*, 108(11):1812–1833.

Stephen Stark, Oleksandr S. Chernyshenko, Fritz Drasgow, Christopher D. Nye, Leonard A. White, Tonia Heffner, and William L. Farmer. 2014. From ABLE to TAPAS: A new generation of personality tests to support military selection and classification decisions. *Military Psychology*, 26(3):153–164.

Emiel van Miltenburg, Chris van der Lee, and Emiel Krahmer. 2021. Preregistering NLP research. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 613–623, Online. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623.

A Collection of Social Desirability Ratings from Human Raters

The human-reference desirability ratings used in Stages I and II came from a pre-existing human-rating dataset collected by the authors in July 2025 using an online survey with participants recruited through Prolific. The dataset was originally collected for a separate research purpose before the preregistration of the present study and has not been reported in a separate publication. The preregistered protocol for the present study specified how this fixed dataset would be used as the human-reference benchmark for Stage I and as the human-rating input for Stage II scale construction.

The dataset contained social desirability ratings for two English HEXACO-based item pools: 60

sentence-format items from the HEXACO-60 inventory, used under the research-use conditions stated on the HEXACO-PI-R materials website,[3] and 60 adjective-format items from the English HEXACO Adjective Scales, used as publicly available published research materials (Romano et al., 2023).

Participants completed one scale-format condition only and rated all 60 items in their assigned condition. The final human-rating dataset used in the present analyses included 277 participants in the sentence-format condition and 321 participants in the adjective-format condition.

A.1 Use of the Pre-existing Dataset in the Preregistered Pipeline

The original human-rating data collection was completed before the preregistration of the present study and was not itself part of the preregistered data-collection procedures. In the preregistered protocol, however, this pre-existing dataset was treated as a fixed human-reference input.

The fixed item-level mean human ratings were used in two parts of the present validation pipeline. First, they served as the human-reference benchmark for evaluating LLM-generated social desirability ratings in Stage I. Second, they served as the human-rating input for constructing the human-based forced-choice inventory in Stage II. No additional human-reference social desirability ratings were collected after preregistration. The LLM-rating collection, forced-choice scale construction, Stage III participant study, scoring model, and confirmatory analyses were conducted according to the preregistered protocol.

A.2 Participant Recruitment and Eligibility

Participants in the original human-rating study were recruited through Prolific and completed the task through an external online survey. The Prolific study was configured to recruit 602 participants. Participants were eligible if they were 18–69 years old, reported English as their first language, and had a Prolific approval rate between 95% and 100%. Participants were required to complete the study on a desktop or laptop computer. Each Prolific participant was allowed to complete the study only once. Participants received £0.70 for completing the study, corresponding to an estimated hourly rate of approximately £8.40 based on the expected survey duration of 5 minutes. Payment was provided through Prolific only for complete and valid responses, as stated on the consent page.

Of the 769 participants who started the survey, 610 submitted data. Of the 610 submitted responses, 12 were excluded: eight because of excessively short completion times (defined as $M-3SD$, approximately < 2 minutes) and four because participants voluntarily withdrew. While 602 participants were approved, technical errors resulted in the loss of four records, yielding a final analytic sample of 598 participants.

A.3 Survey Procedure and Rating Instructions

Before beginning the original human-rating survey, participants were shown an informed-consent page. The text of the consent page is reproduced below.

> You are invited to participate in a research study on psychological questionnaire improvement.
>
> Please read this consent form carefully before agreeing to participate.
>
> **Purpose and Procedures**
>
> This study examines the social desirability of personality trait questions. You will be asked to complete questionnaires about how different personality descriptions are perceived.
>
> The entire study will take approximately 5 minutes to complete.
>
> **Risks and Benefits**
>
> There are no anticipated risks beyond those encountered in daily life. While there may be no direct benefits to you, your participation will help improve psychological measurement tools.
>
> **Voluntary Participation**
>
> Your participation is completely voluntary. You may withdraw at any time without penalty, though compensation requires complete participation.
>
> **Confidentiality**
>
> All data collected in this study will be kept strictly confidential and will be fully anonymized. The aggregated statistical results of this study may be presented at academic conferences or published in academic journals.
>
> Furthermore, the anonymized individual-level data (your responses, with all personally identifying information removed) may be shared with other researchers or made publicly available in a data repository for future research and verification purposes.
>
> **Technical Requirements**
>
> Please use a desktop/laptop computer with a mouse/trackpad (no touchscreens).
>
> Ensure a stable internet connection.
>
> Complete the study in one uninterrupted session.
>
> Maintain your browser window size.

[3] https://hexaco.org/hexaco-inventory

Compensation

Upon completion, you will be redirected to Prolific to receive payment. Payment will only be provided for complete and valid responses. Responses that show clear signs of random clicking or inattentive participation will not be compensated.

Contact Information

Principal Investigator: Kyosuke Bunji (Kobe University)

Email: bunji@bear.kobe-u.ac.jp

This research was approved by Human Ethics Committee of Graduate School of Business Administration, Kobe University.

I consent to participate in this study

By clicking the button above, you acknowledge that you have read and understood this consent form and voluntarily agree to participate in this research study.

After providing informed consent, participants were shown a brief explanation of socially desirable responding in personality assessment contexts. The explanation emphasized that the task concerned how endorsing a personality description would be perceived by others, rather than whether the description actually applied to the participant.

Participants were randomly assigned to either the sentence-format or adjective-format condition.

The rating page displayed the following format-specific instruction. The bracketed alternatives indicate the wording used in the sentence-format and adjective-format conditions, respectively, and were not displayed to participants.

> How would answering "Yes" to the following personality [statements/traits] be perceived by others?
>
> Please rate each [statement/trait] on a 7-point scale from -3 (Very Unfavorable) to $+3$ (Very Favorable).
>
> Please respond to all items based on your initial impression. If you are uncertain about an item, please select 0 (Neutral).

In both conditions, participants rated all 60 items in their assigned scale-format condition using the same 7-point response scale:

$$-3, -2, -1, 0, +1, +2, +3,$$

where -3 indicated "Very Unfavorable," $+3$ indicated "Very Favorable," and 0 indicated "Neutral."

The item order was randomized within each condition.

A.4 Sample Sizes and Data Retention

The retained human-rating dataset contained 598 participants in total: 277 participants in the sentence-format condition and 321 participants in the adjective-format condition. All retained participants provided ratings for the 60 items in their assigned condition.

Participants in the retained human-rating dataset had a mean age of 40.77 years ($\mathrm{SD} = 12.93$). The gender distribution was 329 males, 265 females, and four individuals who preferred not to disclose their gender. The majority of participants were from the United States ($n = 324$), followed by the United Kingdom ($n = 108$), Canada ($n = 62$), Australia ($n = 31$), and other countries ($n = 73$).

A.5 Response Coding and Aggregation

For the present study, item-level mean human ratings were computed for each item and scale format. The resulting item-level mean human ratings were treated as fixed human-reference desirability estimates throughout the present study. These estimates were used as the human-reference benchmark in Stage I and as the human-rating input for Stage II scale construction.

B Collection of Social Desirability Ratings from LLMs

B.1 Overview of the Rating Procedure

In Stage I, large language models (LLMs) were used to generate social desirability ratings for personality items derived from the HEXACO-60 and the HEXACO Adjective Scales. Ratings were collected separately for sentence-format items and adjective-format items. Three LLMs were included in the analyses: GPT-5.4 (OpenAI), Gemini 2.5 Pro (Google), and Claude Opus 4.6 (Anthropic).

For each item, the same prompt was repeatedly submitted to the corresponding LLM. Following the preregistered protocol, each item was evaluated 15 times for each model and scale format. Ratings were collected using a fixed prompt that mirrored the human-rating instruction and used the same seven-point -3 to $+3$ response scale.

Each API request contained a single personality item, and responses were collected independently across repeated queries. Item presentation order was shuffled separately for each repetition cycle. Raw model outputs were stored prior to parsing and aggregation.

The repeated ratings were aggregated by computing the mean rating for each item and model. These aggregated estimates were then used in the analyses reported in Stage I and in the scale construction procedure described in Stage II.

B.2 Prompt Design

The following prompt template was used:

```
You will be shown one [sentence/adjective] that
describes a personality trait.
For this [sentence/adjective], please rate how
answering "Yes" would be perceived by others.
Please rate this [sentence/adjective] on a
7-point scale from -3 (Very Unfavorable) to +3
(Very Favorable), i.e., -3, -2, -1, 0, 1, 2, and
3.
Please respond to this item based on your initial
impression.
If you are uncertain about this item, please
select 0 (Neutral).
Please return EXACTLY ONE integer.
```

Here, [sentence/adjective] was replaced with the appropriate term depending on the scale format.

For each request, the target item was inserted into the prompt as a single stimulus item. Responses were collected as integer ratings and aggregated across repeated queries for subsequent analyses.

B.3 Response Processing and Aggregation

Because LLM outputs can vary across repeated queries, ratings were collected multiple times for each item. Specifically, 15 repeated responses were obtained for each combination of item, model, and scale format.

Returned responses were processed automatically and checked for validity. Responses were considered valid only when the output consisted of a single integer value between -3 and $+3$. Responses failing to satisfy this criterion were flagged as invalid.

Following the preregistered protocol, invalid responses were resubmitted to the same model using the identical prompt. Up to three retry rounds were performed for invalid responses. Valid retry responses replaced the corresponding invalid outputs prior to aggregation.

For each item, the repeated valid ratings were aggregated by computing their mean value. These mean ratings served as the final LLM-based desirability estimates used throughout the subsequent analyses.

B.4 Model and API Settings

Following the preregistered protocol, the study used the most recent high-capacity models that were stably available through publicly accessible APIs at the time of data collection. The evaluated model families were OpenAI's GPT series, Google's Gemini series, and Anthropic's Claude series. Specifically, ratings were collected using GPT-5.4 (OpenAI), Gemini 2.5 Pro (Google), and Claude Opus 4.6 (Anthropic). Table 4 summarizes the model identifiers and snapshots used in the experiments reported in this paper. LLM-based desirability ratings were collected between April 9 and April 10, 2026.

Label in text	Model identifier / snapshot
GPT-5.4	gpt-5.4-2026-03-05
Gemini 2.5 Pro	gemini-2.5-pro
Claude Opus 4.6	claude-opus-4-6

Table 4: Model identifiers and snapshots used in the experiments reported in this paper.

OpenAI and Anthropic models were accessed from R via the `ellmer` package, whereas Gemini models were accessed via the Google Gemini REST API. Additional API communication and batch-processing procedures were implemented using the `httr2` and `jsonlite` packages. Only models with stable API availability and explicitly stated API pricing were included. LLM ratings were collected through commercially accessible provider APIs under the standard API access terms in effect at the time of data collection; no provider-side model weights, non-public training data, or API-internal system information were accessed or redistributed.

All generation parameters (e.g., temperature and top-p) used the default settings of the corresponding provider APIs. No operations assigning specific personas, roles, or identities to the LLMs were performed. No hyperparameter tuning or prompt optimization was performed. Because commercial APIs may change over time and some providers do not guarantee deterministic generation under default settings, exact replication of raw LLM outputs may not always be possible even when using the same model identifiers. In addition, system prompts and user prompts were not separated; all instructions and item presentations were provided within a single user prompt using a common

prompting procedure across models and scale formats.

Because the evaluated LLMs were accessed through commercial APIs, full parameter counts and provider-side infrastructure details, including hardware configuration and GPU usage, were not publicly disclosed or otherwise available to the authors at the time of data collection. Accordingly, we report provider names, model identifiers/API snapshots, access dates, generation settings, and response-processing procedures rather than parameter counts or hardware-level compute details.

C Construction of the Graded Forced-Choice Inventory

C.1 Item Pool and Desirability Inputs

The graded forced-choice (GFC) inventories constructed in Stage II were based on the HEXACO-60 sentence items and the HEXACO Adjective Scales. Scale construction was conducted separately for sentence-format and adjective-format items and independently for the human-based and LLM-based desirability ratings.

Each item was associated with a HEXACO domain label (Honesty–Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, or Openness to Experience) and a keying direction indicating whether the item was positively or negatively keyed. Item-level social desirability values were defined as the mean desirability ratings obtained in Stage I for each source and scale format.

Prior to optimization, item exclusion rules preregistered in Stage I were applied consistently across all sources and scale formats. Items with incomplete LLM desirability estimates were excluded from the optimization procedure.

C.2 Generation of Candidate Pairs

Candidate forced-choice pairs were generated separately for each combination of rating source and scale format. Only cross-domain item pairs were considered; that is, pairs composed of two items belonging to different HEXACO domains.

For each candidate pair, the absolute difference in desirability ratings between the two items was computed:

$$\Delta_{jj'} = |s_j - s_{j'}|,$$

where s_j and $s_{j'}$ denote the desirability estimates for items j and j'.

In addition, a mixed-key indicator was defined for each candidate pair to indicate whether the two items differed in keying direction:

$$g_{jj'} = \mathbb{I}[k_j \neq k_{j'}],$$

where $k_j \in \{-1, +1\}$ represents the keying direction of item j.

Candidate pairs were ordered according to ascending item IDs. This ordering was retained throughout the optimization procedure and subsequently carried forward to the Stage III analyses.

C.3 Optimization-Based Pair Selection

For each source and scale format, exactly 21 forced-choice item pairs were selected using a multi-stage lexicographic mixed-integer optimization procedure implemented in `R` (R Core Team, 2025) using the `ompr` and `ROI` frameworks. Optimization was performed using the Gurobi solver (Gurobi Optimization, LLC, 2026).

The optimization simultaneously minimized desirability mismatch within item pairs while enforcing multiple structural balance constraints.

C.3.1 Stage 1: Minimax Optimization

In the first optimization stage, the objective was to minimize the maximum desirability difference across all selected item pairs. Let $x_{jj'} \in \{0, 1\}$ indicate whether candidate pair (j, j') was selected, and let m denote the maximum within-pair desirability difference. The optimization objective was:

$$\min_{x,m} \quad m$$

subject to the following constraints:

$$\sum_{(j,j')} x_{jj'} = 21, \tag{1}$$

$$\Delta_{jj'} x_{jj'} \leq m \qquad \forall (j, j'), \tag{2}$$

$$\sum_{(j,j'):j=\ell \text{ or } j'=\ell} x_{jj'} \leq 1 \qquad \forall \ell. \tag{3}$$

The first constraint ensured that exactly 21 item pairs were selected. The second implemented the minimax criterion. The third enforced item uniqueness such that each item could appear in at most one pair.

C.3.2 Constraint Conditions

Several additional structural constraints were imposed during optimization.

First, mixed-key pairs were constrained to constitute between 40% and 60% of the selected item

pairs:

$$0.40P \leq \sum_{(j,j')} g_{jj'} x_{jj'} \leq 0.60P,$$

where $P = 21$ denotes the number of selected pairs.

Second, balanced domain representation was enforced. Because 21 pairs contain 42 selected items, each HEXACO domain was constrained to appear exactly 7 times in the final inventory.

Third, balance across HEXACO domain combinations was imposed. Because there are 15 possible unordered HEXACO domain combinations, each domain-pair type was constrained to occur between 1 and 2 times, representing a balanced feasible allocation for the 21 selected pairs.

Finally, within each HEXACO domain, positive- and negative-keyed items were constrained such that each sign accounted for at least 30% of the selected items within that domain.

C.3.3 Stage 2: Minimization of Total Squared Desirability Differences

Let $m^\star$ denote the optimal minimax value obtained in Stage 1. In the second optimization stage, the minimax solution was retained by constraining:

$$m \leq m^\star + \varepsilon,$$

where $\varepsilon = 10^{-6}$.

Within this feasible region, the optimization minimized the total squared desirability mismatch across all selected pairs:

$$\min_x \sum_{(j,j')} (s_j - s_{j'})^2 x_{jj'}.$$

This stage selected the solution with the smallest overall desirability discrepancy while preserving the Stage 1 minimax optimum.

C.3.4 Stage 3: Rank-Sum Tie-Breaking

Because multiple Stage-2-optimal solutions could exist, an additional optimization stage was applied to obtain a deterministic solution among equivalent optima.

Candidate pairs were assigned deterministic ranks according to ascending item IDs. Among all solutions achieving the Stage 2 optimum within a numerical tolerance, the optimization minimized the sum of candidate ranks across selected pairs:

$$\min_x \sum_{(j,j')} r_{jj'} x_{jj'},$$

where $r_{jj'}$ denotes the deterministic rank assigned to candidate pair (j, j').

This procedure implemented the preregistered rank-sum tie-breaking rule while preserving the Stage 2 optimum.

C.3.5 Stage 4: Residual Lexicographic Tie-Breaking

After the rank-sum optimization, the solution space was checked for remaining equivalent solutions. If residual ties still existed, a final deterministic lexicographic tie-breaking procedure was applied.

This procedure sequentially fixed candidate-pair selection decisions in ascending candidate order while preserving all previous optimization criteria and constraints. The resulting solution therefore represented a unique canonical solution among all remaining equivalent optima.

The full optimization pipeline consisted of the following sequential stages:

1. Stage 1 minimax optimization
2. Stage 2 sum-of-squares optimization
3. Stage 3 rank-sum tie-break
4. Stage 4 residual lexicographic tie-break (if necessary)

C.4 Output Formatting and Scale Construction

The final optimized solutions were exported separately for each source and scale format. For each selected pair, the exported information included item identifiers, item texts, HEXACO domains, keying directions, desirability ratings, and within-pair desirability differences.

The resulting human-based and LLM-based GFC inventories were subsequently administered in Stage III for the estimation and comparison of latent trait scores.

D Full TIRT model specifications

D.1 Joint Pair Bank and Response Data

Latent trait estimation in Stage III was conducted using a graded Thurstonian item response theory (TIRT) model applied to the forced-choice responses obtained from the human-based and LLM-based inventories.

Analyses were conducted separately for sentence-format and adjective-format inventories. For each scale format, a joint pair bank was

constructed by combining the item pairs from the corresponding human-based and LLM-based inventories. Because each inventory consisted of 21 forced-choice pairs, the joint pair bank contained up to 42 pair instances per scale format.

Although some items overlapped across inventories, items originating from different scale sources were treated as distinct item instances during calibration. Each forced-choice response was recorded using a 7-category ordinal response scale ranging from 1 to 7.

Participants completed both the human-based and LLM-based inventories within the same scale format condition. Responses to the two inventories were jointly calibrated to place latent trait estimates on a common metric.

D.2 Graded Thurstonian IRT Model

Let Y_{np} denote the response of participant n to pair p, where:

$$Y_{np} \in \{1, 2, \ldots, 7\}.$$

Responses were modeled using an ordered logistic formulation:

$$Y_{np} \sim \mathrm{OrderedLogistic}(\eta_{np}, \mathbf{c}_p),$$

where η_{np} denotes the latent utility difference for pair p, and $\mathbf{c}_p$ represents the ordered category thresholds associated with pair p.

The latent utility difference was defined as:

$$\eta_{np} = \frac{1}{\sqrt{2}} \left(a_{R(p)} z_{nR(p)} - a_{L(p)} z_{nL(p)} \right),$$

where:

- $a_{R(p)}$ and $a_{L(p)}$ denote the discrimination parameters for the right and left items in pair p,
- $z_{nR(p)}$ and $z_{nL(p)}$ denote the latent utilities associated with the corresponding items,
- the scaling factor $1/\sqrt{2}$ adjusts for variance inflation induced by differencing latent utilities.

The latent utility for item j was defined as:

$$z_{nj} = \boldsymbol{\theta}_n^\top \mathbf{q}_j,$$

where:

- $\boldsymbol{\theta}_n$ is the six-dimensional latent trait vector for participant n,
- $\mathbf{q}_j$ is the binary domain-loading vector for item j.

Each item loaded on a single HEXACO domain.

Item discrimination parameters were parameterized as:

$$a_j = g_j a_j^+,$$

where:

- $a_j^+ > 0$ denotes the positive discrimination magnitude,
- $g_j \in \{-1, +1\}$ represents the item keying direction.

Each pair was associated with six ordered threshold parameters:

$$\mathbf{c}_p = (c_{p1}, \ldots, c_{p6}),$$

where c_{pk} denotes the k-th threshold for pair p, with $k = 1, \ldots, 6$; these six thresholds separate the seven ordered response categories.

D.3 Prior Distributions

Participant latent traits were modeled as:

$$\boldsymbol{\theta}_n = \mathbf{L}_\Omega \boldsymbol{\theta}_n^{\mathrm{raw}},$$

where:

$$\boldsymbol{\theta}_n^{\mathrm{raw}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

The latent trait correlation matrix Ω was parameterized through its Cholesky factor L_Ω, such that

$$\Omega = L_\Omega L_\Omega^\top.$$

The Cholesky factor was assigned an LKJ Cholesky prior:

$$L_\Omega \sim \mathrm{LKJCholesky}(2).$$

Positive discrimination magnitudes were assigned half-normal priors:

$$a_j^+ \sim \mathcal{N}^+(0, 0.5).$$

Threshold parameters were assigned normal priors:

$$c_{pk} \sim \mathcal{N}(0, 1.5).$$

The second argument specifies the normal distribution's standard deviation.

D.4 Joint Calibration and Form-Specific Scoring

For each scale format, human-based and LLM-based inventories were first jointly calibrated using the combined pair bank. This procedure jointly estimated item discrimination parameters for form-specific item instances and pair-specific threshold parameters across both inventories, together with a shared latent trait correlation structure.

After joint calibration, form-specific participant scoring was conducted separately for the human-based and LLM-based inventories. During this scoring stage, the item discrimination parameters, threshold parameters, and latent trait correlation structure estimated from the joint calibration model were fixed.

Participant latent trait estimates were then obtained separately for:

- the human-based inventory,
- the LLM-based inventory.

This procedure ensured that latent trait estimates derived from the two inventories were placed on a common metric prior to the correlation analyses reported in Stage III.

Participants with fewer than 10 observed pair responses within a given inventory would have been excluded from the corresponding scoring analyses.

D.5 Estimation Settings

All TIRT models were estimated in Stan using Hamiltonian Monte Carlo sampling via `cmdstanr` (Carpenter et al., 2017; Gabry et al., 2024). Estimation was performed separately for sentence-format and adjective-format inventories.

For all models, four Markov chains were used with:

- 1000 warmup iterations,
- 1000 post-warmup sampling iterations per chain,
- `adapt_delta` set to 0.95,
- `max_treedepth` set to 12.

Sampling reproducibility was facilitated through fixed random seeds.

Model convergence was evaluated using $\hat{R}$ statistics and inspection of divergent transitions. The resulting convergence summaries are reported in the main Results section.

E Details of Human Data Screening and Sample Sizes

E.1 Participant Recruitment and Survey Administration

Stage III data were collected using an online survey platform in combination with participant recruitment through Prolific. Participants were eligible to participate if they were native English speakers between 18 and 69 years of age. Participants were additionally required to complete the study using a desktop or laptop computer.

Participants were recruited through Prolific and received £1.35 for participation, corresponding to an estimated hourly rate of approximately £9 based on the expected survey duration of 9 minutes.

Participants completed either the sentence-format or adjective-format version of the graded forced-choice (GFC) inventory. Within each scale-format condition, the presentation order of the human-based and LLM-based scales was counterbalanced across participants.

Raw survey-response files and participant-tracking information were processed using the Stage III preprocessing pipeline. Screening variables, exclusion indicators, and participant-level tracking summaries were generated automatically during preprocessing.

E.2 Survey Procedure and Rating Instructions

Before beginning the Stage III forced-choice survey, participants were shown an informed-consent page. The text of the consent page is reproduced below.

> You are invited to participate in a research study on psychological questionnaire improvement.
>
> Please read this consent form carefully before agreeing to participate.
>
> **Purpose and Procedures**
>
> In this study, you will complete two brief questionnaires about your personality. The purpose is to improve psychological measurement tools.
>
> This survey will take approximately 9 minutes to complete.
>
> **Risks and Benefits**
>
> There are no anticipated risks beyond those encountered in daily life.
>
> While there may be no direct benefits to you, your participation will help improve psychological measurement tools.
>
> **Voluntary Participation & Withdrawal**
>
> Your participation is voluntary. If you feel uncomfortable at any point during the survey, you

may stop responding without penalty. However, in order to receive payment, you must provide informed consent and complete the entire survey.

Confidentiality

All data collected in this study will be kept strictly confidential and will be fully anonymized. The aggregated statistical results may be presented at academic conferences or published in academic journals.

Furthermore, the anonymized individual-level data (your responses, with all personally identifying information removed) may be shared with other researchers or made publicly available in a data repository for future research and verification purposes.

Technical Requirements

Desktop/Laptop computer only (Mobile/Tablet not supported).

Ensure a stable internet connection.

Do not reload the browser during the task.

Compensation

Upon completion, you will be redirected to Prolific to receive payment. Payment will only be provided for complete and valid responses. Responses that show clear signs of random clicking or inattentive participation will not be compensated.

Contact Information

Principal Investigator: Kensuke Okada (The University of Tokyo)

Email: ken@p.u-tokyo.ac.jp

Yui Furukawa (The University of Tokyo) Kyosuke Bunji (Kobe University)

This research was approved by the Research Ethics Committee of the University of Tokyo.

I consent to participate in this study

By clicking the button above, you acknowledge that you have read and understood this consent form and voluntarily agree to participate in this research study.

After reviewing the informed-consent page, participants completed the assigned forced-choice inventory administered through an online survey platform. Participants were randomly assigned to either the sentence-format or adjective-format condition. Within each condition, participants completed both the human-based and LLM-based versions of the forced-choice inventory, and the presentation order of the two scales was counterbalanced across participants.

Participants received the following instruction before responding to the forced-choice items:

> For each of the following pairs of [sentences/adjectives], please indicate which one describes you more accurately and by how much using a 7-point bipolar scale:

Here, [sentence/adjective] was replaced with the appropriate term depending on the scale format. The response options were presented as follows:

> 1: LEFT [sentence/adjective] describes me much more accurately
> 2: LEFT [sentence/adjective] describes me moderately more accurately
> 3: LEFT [sentence/adjective] describes me slightly more accurately
> 4: About the same
> 5: RIGHT [sentence/adjective] describes me slightly more accurately
> 6: RIGHT [sentence/adjective] describes me moderately more accurately
> 7: RIGHT [sentence/adjective] describes me much more accurately.

Participants additionally received the following instruction:

> We know that, in many cases, the answer depends on the specific situation. Please answer by thinking about how much each [sentence/adjective] describes you "on average" through different situations.

Before beginning the survey, participants were informed that the survey included attention-check items intended to ensure data quality and were instructed to read all questions and instructions carefully.

The survey included two attention-check item pairs, with one pair presented randomly within each inventory block. For sentence-format inventories, the attention-check pairs were:

> Pair 1: "I am currently participating in an online survey." –
> "I have never used the internet or a computer in my entire life."
>
> Pair 2: "I spend all hours of every day asleep and have never been awake." –
> "I am awake at least some of the time."

For adjective-format inventories, the attention-check pairs were:

> Pair 1: human – nonhuman
> Pair 2: nonliving – living

In each pair, one option represented a clearly implausible or contradictory response.

Following Prolific guidelines for attention checks, responses selecting the clearly implausible option were treated as failed attention checks, whereas responses up to and including the neutral option ("About the same") were retained as acceptable responses. Participants who failed both attention checks were excluded from subsequent analyses.

The survey additionally included procedures for detecting incomplete responses, exceptionally short completion times, and invariant responding patterns, as described in the following subsection.

Only participants who provided informed consent after reviewing this page were allowed to proceed to the forced-choice task.

E.3 Sample Sizes and Data Retention

Several preprocessing and validation procedures were applied prior to latent trait estimation analyses. These procedures were designed to identify invalid responses, incomplete administrations, and cases unsuitable for Thurstonian item response theory (TIRT) estimation.

The preprocessing pipeline evaluated:

- invalid or missing participant identifiers,
- survey-design inconsistencies,
- exceptionally short completion times,
- failures on attention-check items,
- straightlining or invariant responding patterns,
- responses outside the expected response range.

Participants meeting one or more exclusion criteria were excluded prior to the confirmatory analyses. In addition, participants with fewer than 10 observed item-pair responses within either scale source would have been excluded prior to latent trait estimation. However, all retained participants provided sufficient observed responses for TIRT estimation.

Raw survey records were available for 404 participants. Of these, 400 participants had approved Prolific submissions and were eligible for analysis. No approved participants met the exclusion criteria. Accordingly, 400 participants were included in the final Stage III analyses. Participants ranged in age from 18 to 69 years ($M = 36.78$, $SD = 12.78$). The sample consisted of 221 men, 178 women, and 1 participant who preferred not to disclose their gender. Final sample sizes by scale-format condition and presentation-order condition are summarized in Table 5.

Presentation order	N
Sentence	
Human-based first	103
LLM-based first	111
Adjective	
Human-based first	88
LLM-based first	98

Table 5: Final sample sizes by scale format and presentation order. Human-based first indicates that participants completed the human-based inventory prior to the LLM-based inventory. LLM-based first indicates the reverse presentation order.

F Analysis Details and Preregistration Mapping

F.1 Preregistration Overview

The study protocol, hypotheses, planned analyses, and optimization procedures were preregistered before LLM rating collection and Stage III data collection. The preregistration document is available at https://osf.io/xca6z/overview.

The preregistration specified the procedures for:

- collection of LLM-based social desirability ratings,
- optimization-based construction of graded forced-choice (GFC) inventories,
- administration of human-based and LLM-based inventories in Stage III,
- estimation of latent trait scores using Thurstonian item response theory (TIRT),
- evaluation of correspondence between human-based and LLM-based outcomes.

The primary analyses reported in the main text followed the preregistered analysis plan. Additional exploratory and implementation-specific analyses are described below.

F.2 Mapping Between Preregistered and Reported Analyses

Table 6 summarizes the correspondence between preregistered analyses and the analyses reported in the main text and Appendix.

Preregistered analysis/component	Reported section(s)
LLM-based desirability rating collection	Methods; Appendix B
Optimization-based GFC scale construction	Methods; Appendix C
Stage I rating-level correlations	Results Section 4.1
Stage II pair-selection analyses and overlap summaries	Results Section 4.2; Appendix H
Stage III latent trait correspondence analyses	Results Section 4.3
TIRT model estimation	Methods; Appendix D
Exploratory supplementary analyses	Appendices G–J

Table 6: Mapping between preregistered analyses and reported results

F.3 Exploratory Analyses

Several analyses reported in the Appendix were conducted for exploratory or supplementary purposes and were not central components of the preregistered primary analyses.

These exploratory analyses included:

- concordance correlation coefficient (CCC) analyses,
- Bayes factor comparisons between sentence and adjective correlations,
- ELoT overlap analyses,
- Human–Human rating-variability benchmarks for Stage I correlations and Stage II pair-level Jaccard overlap,
- overlap breakdown summaries for Stage II scale construction,
- descriptive summaries of within-pair desirability differences,
- optimization-sensitivity analyses, including tie structure, solution multiplicity, cross-evaluation, and z-standardization invariance,
- Bland–Altman analyses of latent trait estimates,
- additional Stage III agreement and score-precision diagnostics, including CCCs, posterior SD summaries, mean differences, and shared-item parameter stability,
- exploratory open-weight model analyses for Llama and Mistral in Stages I and II.

These analyses were conducted to provide additional descriptive information regarding agreement patterns, scale composition, and latent trait correspondence beyond the preregistered primary outcome analyses.

F.4 Additional Technical Clarifications

Although the Stage I human reference ratings and LLM ratings were both originally collected using the same -3 to $+3$ response scale, the archived human-rating dataset stored responses using an internally recoded 0–6 format for data-processing purposes. Because Pearson correlation coefficients are invariant under linear transformation, this coding representation did not affect the primary Stage I correlation analyses reported in the main text. For visualization and exploratory concordance correlation coefficient (CCC) analyses, human ratings were transformed back to the original -3 to $+3$ response metric.

F.5 Software and Computational Environment

All analyses were conducted in `R` version 4.5.2. Mixed-integer optimization procedures were implemented with the Gurobi solver. Bayesian TIRT models were estimated in Stan using `cmdstanr`.

Random seeds were fixed throughout data collection, optimization, and Bayesian estimation procedures to facilitate reproducibility.

F.6 Deviations from the preregistration protocol

Several implementation-related deviations and clarifications from the preregistered protocol should be noted.

Adjective-item mismatch and item exclusion. The preregistration specified that all 60 HEXACO adjective items would be included in the confirmatory analyses. After data collection, however, we identified a discrepancy in Item 44 between the

English adjective-scale materials linked from the HEXACO Personality Inventory–Revised website and the published HAS article. Both the human-reference ratings and the LLM ratings in the present study were collected using the website-distributed version, which lists Item 44 as *sicure* (Romano and Perugini, 2023, p. 12). By contrast, the appendix of the published HAS article lists the corresponding item as the Italian–English pair *Sicuro–Secure* (Romano et al., 2023, p. 2529). Because this source-material discrepancy was detected after data collection, we excluded Item 44 from the confirmatory adjective-format analyses. As a result, the adjective-format analyses used 59 items rather than the 60 specified in the preregistration.

Although LLM ratings for Item 44 were collected as originally planned, the item was excluded because of the discrepancy with the published item wording before the Stage I correlation analyses and Stage II optimization-based scale construction. Consequently, adjective-format confirmatory analyses used 59 items instead of the preregistered 60. This exclusion was based on item-content consistency and was determined independently of the study outcomes.

Additional deterministic tie-breaking in Stage II optimization. The preregistered Stage II optimization procedure was retained as registered through the deterministic tie-break based on minimizing a linear function of candidate-pair indices under the predefined candidate ordering. However, during implementation, we identified that this linear tie-break could still yield multiple equally optimal solutions because different selected-pair sets could share the same index-sum value.

To resolve only these residual ties, we added one final deterministic tie-breaking step after the preregistered linear tie-break. Specifically, within the set of solutions tied on all preregistered criteria, candidate pairs were traversed sequentially according to the same predefined candidate ordering and fixed to 1 whenever at least one feasible completion remained; otherwise they were fixed to 0. This procedure yielded a unique canonical solution.

This refinement did not alter the candidate pool, optimization constraints, the first-stage minimax objective, the second-stage sum-of-squared desirability-difference objective, or the preregistered fixed-order linear tie-break. The additional tie-breaking step was applied identically to the human-rating and LLM-rating optimization runs and served only to ensure deterministic uniqueness and reproducibility. Because the Stage II analyses were preregistered as descriptive rather than confirmatory, this modification is treated as a procedural clarification rather than a change in substantive inferential criteria.

Clarification regarding attention-check wording. The preregistration contained inconsistent wording regarding the placement of the attention checks used in Stage III. In the *Sampling – Data collection procedures* section, the preregistration correctly stated that submissions failing both attention checks introduced during the survey could be rejected according to Prolific's guidelines. However, in the *Analysis Plan – Data inclusion and exclusion* section, the preregistration mistakenly referred to "both attention checks introduced at the beginning of the survey."

This latter wording was an editing error and should not be interpreted as restricting the attention checks to the beginning of the survey. The exclusion criterion applied in the study was failure of both attention checks regardless of their position within the survey. This clarification does not affect the substantive analyses, exclusion logic, or preregistered inference criteria.

G Results of the Exploratory Analyses in Stage I

G.1 Descriptive Summaries of Rating Distributions

Table 7 provides descriptive summaries of the item-level mean human-reference and LLM ratings used in Stage I. As reported in the main text, Pearson correlations between human-reference and LLM ratings were very high across models and scale formats, indicating strong linear correspondence in the relative ordering of item desirability. At the same time, the descriptive summaries show that human raters and LLMs did not use the seven-point response scale in the same way.

For sentence-format items, the observed range of item-level mean human ratings was 4.148, whereas all three LLMs spanned the full possible range of 6.000. The standard deviation of human item-level means was also smaller (SD = 1.277) than those of the LLM item-level means (SDs = 2.011–2.303). The same pattern was observed for adjective-format items: the range of human item-level means was 4.969, whereas all three LLMs again spanned the

Scale format	Source	N	M	SD	Min	Max	Range
Sentence	Human	60	0.077	1.277	-2.032	2.116	4.148
	GPT	60	0.128	2.022	-3.000	3.000	6.000
	Gemini	60	0.102	2.303	-3.000	3.000	6.000
	Claude	60	0.104	2.011	-3.000	3.000	6.000
Adjective	Human	59	0.116	1.720	-2.508	2.461	4.969
	GPT	59	0.277	2.417	-3.000	3.000	6.000
	Gemini	59	0.238	2.488	-3.000	3.000	6.000
	Claude	59	0.228	2.374	-3.000	3.000	6.000

Table 7: Descriptive summaries of item-level social desirability ratings. Values are computed from item-level mean ratings after applying the Stage I item-exclusion rule. All ratings are expressed on the original -3 to $+3$ response metric. Range = Max $-$ Min. These summaries are descriptive and were not part of the preregistered Stage I support criterion.

full possible range of 6.000, and the human standard deviation (SD = 1.720) was smaller than the LLM standard deviations (SDs = 2.374–2.488). Mean ratings also tended to be slightly higher for LLMs than for human raters, particularly in the adjective format.

These results indicate that the high Stage I correlations should be interpreted as evidence of strong linear correspondence rather than identical absolute scale use. In other words, LLM ratings closely tracked human ratings in terms of relative item desirability, but the LLMs showed systematically broader dispersion and more extreme use of the response metric. This distinction motivates the supplementary concordance analyses reported below, which evaluate agreement while taking absolute scale use into account.

G.2 Agreement Based on Concordance Correlation Coefficients

As shown in Table 8, CCCs between human and LLM ratings ranged from 0.804 to 0.871 for sentence items and from 0.915 to 0.928 for adjective items across models.

These values were slightly lower than the corresponding Pearson correlation coefficients reported in the main analysis, reflecting the fact that the CCC incorporates both precision and absolute agreement and is therefore sensitive to differences in both correlation strength and absolute scale agreement. Consistent with the Pearson correlation results, CCC values were higher for adjective items than for sentence items, indicating somewhat higher agreement for adjective-format items across models. Overall, the results suggest a high level of concordance between human and LLM desirability ratings when evaluated on a common response scale.

Model	CCC	95% CI	N
Sentence			
GPT	0.850	[0.803, 0.886]	60
Gemini	0.804	[0.759, 0.842]	60
Claude	0.871	[0.837, 0.899]	60
Adjective			
GPT	0.923	[0.901, 0.941]	59
Gemini	0.915	[0.893, 0.933]	59
Claude	0.928	[0.906, 0.946]	59

Table 8: Concordance correlation coefficients between human and LLM ratings. CCC = concordance correlation coefficient. Values indicate agreement between human mean ratings and LLM mean ratings at the item level. 95% confidence intervals are reported in brackets. N denotes the number of items included in each condition.

G.3 Bayes Factor Analyses

For each model, Bayes factors comparing the equality of two correlations were computed with `cor_test()` and `BF()` from the `BFpack` package (Mulder et al., 2021). The analyses used the default joint-uniform prior on the group-specific correlation matrices and compared the equality-constrained model with its complement. Reported values were obtained with 30,000 MCMC iterations and a burn-in of 5,000 iterations.

Model	Sentence r	Adjective r	Difference	BF equal	BF different
GPT	0.942	0.980	-0.039	0.079	12.695
Gemini	0.948	0.980	-0.032	0.165	6.048
Claude	0.963	0.979	-0.015	2.225	0.449

Table 9: Comparison of correlations between sentence and adjective items. Values represent Pearson correlation coefficients between human and LLM mean ratings for sentence and adjective items. Difference indicates the subtraction (sentence – adjective). Differences were computed from unrounded correlations. BF equal denotes the Bayes factor in favor of equal correlations, and BF different denotes the Bayes factor in favor of unequal correlations.

G.4 Comparison Between Sentence and Adjective Items

Table 9 presents a comparison of correlations between sentence and adjective items. Bayes factor analyses suggested evidence for differences in correlations for GPT (strong evidence) and Gemini (moderate evidence), whereas for Claude the evidence was anecdotal and slightly favored the model assuming equal correlations.

However, the differences in correlation coefficients between sentence and adjective items were small across models. The largest observed difference in correlation coefficients was 0.039 (GPT), indicating that the correlations remained high for both sentence and adjective items and were largely comparable in magnitude.

G.5 Effects of ELoT Overlap

ELoT refers to the English List of Traits (Britz et al., 2023). We classified the 59 adjective items included in the analyses according to whether they overlapped with ELoT: 29 items overlapped with ELoT and 30 did not.

Table 10 shows correlations for ELoT overlap and non-overlap items, along with their differences and Bayes factor comparisons. Across all models, correlations were very similar between the two groups, with minimal differences in point estimates.

Consistent with this pattern, the Bayes factor analyses did not provide substantial evidence for differences in correlations between overlap and non-overlap items. The observed differences in correlation coefficients were negligible (ranging from 0.002 to 0.021), suggesting that ELoT overlap was not associated with meaningful differences in the strength of correspondence between human and LLM ratings.

H Results of the Exploratory Analyses in Stage II

As shown in Table 11, the number of items shared between the human- and LLM-based scales ranged from 28 to 33 out of 42 across models and scale formats. At the same time, each scale contained 9 to 14 items not included in the other scale, indicating that the overlap in selected items was partial rather than complete.

At the pair level, overlap between the two approaches was limited. The number of shared item pairs ranged from 1 to 5 for sentence items and from 0 to 1 for adjective items. In particular, for adjective items, almost no item pairs were shared between the two scales. This pattern suggests that while similar items tend to be selected, the specific pairing of items within the constructed scales differs.

Table 12 summarizes within-pair differences in desirability ratings. For LLM-based scales, mean differences ranged approximately from 0.23 to 0.46 for sentence items and from 0.25 to 0.32 for adjective items, depending on the model. For the human-based scale, mean differences were 0.259 for sentence items and 0.169 for adjective items. These values indicate that the desirability differences within selected item pairs were of a similar order of magnitude across scales. Differences tended to be larger for sentence items than for adjective items, although this pattern was not uniform across all models.

Overall, the results suggest that the constructed scales do not contain item pairs with systematically large differences in desirability. This indicates that variation in desirability within pairs is unlikely to be a major factor driving potential discrepancies in latent trait estimates in subsequent analyses.

Model	ELoT overlap r	Non-overlap r	Difference	BF equal	BF different
GPT	0.985	0.964	0.021	1.036	0.965
Gemini	0.977	0.974	0.004	3.969	0.252
Claude	0.974	0.973	0.002	4.068	0.246

Table 10: Comparison of correlations between ELoT overlap and non-overlap items. ELoT refers to the English List of Traits; the overlap and non-overlap groups contained 29 and 30 items, respectively. Values represent Pearson correlation coefficients for ELoT overlap and non-overlap items. Difference indicates the subtraction (overlap – non-overlap). Differences were computed from unrounded correlations. BF equal denotes the Bayes factor in favor of equal correlations, and BF different denotes the Bayes factor in favor of unequal correlations.

	Item		Pair	
Model	Both	Unique	Both	Unique
Sentence				
GPT	29	13	1	20
Gemini	33	9	2	19
Claude	33	9	5	16
Adjective				
GPT	30	12	1	20
Gemini	31	11	0	21
Claude	28	14	0	21

Table 11: Breakdown of item and pair overlap between human- and LLM-based scales. Both denotes the number of items or item pairs included in both the human- and LLM-based scales. Unique denotes the number of elements unique to each individual scale. Because the two scales have equal sizes, this count is the same for both. Results are shown separately for item-level and pair-level selections.

Source	N pairs	Mean	SD
Sentence			
Human	21	0.259	0.353
GPT	21	0.229	0.482
Gemini	21	0.457	0.801
Claude	21	0.229	0.597
Adjective			
Human	21	0.169	0.290
GPT	21	0.257	0.573
Gemini	21	0.324	0.741
Claude	21	0.254	0.649

Table 12: Summary statistics of within-pair differences in desirability ratings. Values summarize absolute differences in desirability ratings between paired items within each selected pair. Statistics are reported separately for human-based and LLM-based scales. Mean and standard deviation (SD) are reported. N pairs denotes the number of item pairs included in each condition.

H.1 Human–Human Rating-Variability Benchmark

For the Stage I rating-correlation benchmark, two human panels of $K = 15$ raters were generated separately for each item format by resampling participants with replacement; each participant's complete set of item ratings within a format was retained as a unit. LLM ratings were resampled from the 15 repeated rating vectors. Within each bootstrap repetition, the same two human panels were used across the three LLMs. Human–LLM correlations were computed against each of the two human panels and averaged on the Fisher-z scale, and the comparison statistic was $\Delta z = z_{\mathrm{HL}} - z_{\mathrm{HH}}$. Percentile bootstrap intervals were computed across 5,000 repetitions. For the Stage II pair-level benchmark, two human panels were generated in the same way, Stage II optimization was rerun independently for each panel using the same structural constraints, objective hierarchy, tolerances, and Gurobi solver as in the main analysis, and pair-level Jaccard overlap was computed between the two human-derived forms. The observed LLM-based forms were held fixed, and the observed Human–LLM pair-level Jaccard values were compared with this Human–Human distribution. This Stage II benchmark used 500 repetitions, with no optimization failures; pair identity was defined by item format and unordered item IDs.

As an exploratory benchmark, we compared the observed human–model discrepancies with the variability expected from resampling human raters. This analysis was intended to contextualize the magnitude of human–model differences relative to Human–Human rating variability, rather than to provide a formal equivalence or non-inferiority test.

For Stage I rating correlations, bootstrap intervals for the human–model minus Human–Human Fisher-z differences included zero in all six model-

by-format conditions. For Stage II pair-level agreement, the mean Human–Human pair-level Jaccard values were 0.027 for sentence-format items and 0.045 for adjective-format items. In all six human–model conditions, the observed pair-level Jaccard values were not below the 2.5th percentile of the corresponding Human–Human bootstrap distribution.

These results suggest that, under these exploratory benchmarks, the observed human–model differences were not clearly larger than the variability induced by resampling human rating panels. However, because this analysis was exploratory and descriptive, it should be read as a comparison with human-rater variability, not as a formal test of equivalence.

H.2 Optimization Sensitivity and Multiple Optimal Solutions

As an exploratory analysis, we examined whether the low exact-pair overlap between human- and LLM-based forms could be partly explained by multiplicity in the optimization landscape and rating-source-specific differences, rather than by a failure of the optimization procedure. Several supplementary checks were conducted, including analyses of zero-gap candidate pairs, the number of alternative optimal solutions, cross-evaluation under the alternative rating source, and invariance under z-standardization and linear rescaling of the LLM desirability ratings.

A zero-gap candidate was an eligible cross-domain pair whose two items had identical mean desirability ratings under the relevant rating source, with equality evaluated to within 10^{-12}.

The optimization landscape contained many zero-gap candidate pairs for the two LLM conditions carried forward to Stage III. The Claude sentence-format condition contained 231 zero-gap candidate pairs, of which 18 appeared in the selected form, whereas the corresponding human sentence-format condition contained 1 zero-gap candidate pair, which was not selected. The GPT adjective-format condition contained 183 zero-gap candidate pairs, of which 14 appeared in the selected form, whereas the corresponding human adjective-format condition contained 2 zero-gap candidate pairs, of which 1 was selected. These zero-gap candidate counts describe eligible item pairs with identical mean desirability ratings and are distinct from the number of alternative optimal pair sets.

For the two LLM conditions carried forward to Stage III, distinct pair sets attaining the same minimax and conditional sum-of-squares optima were enumerated up to a prespecified cap of 200. Enumeration reached this cap in both cases, indicating that at least 200 such solutions existed. By contrast, the human-rating optimization identified one optimal solution for the sentence format and two for the adjective format.

The observed solutions were also cross-evaluated under the alternative rating source. When LLM-selected forms were evaluated using the human-rating objective, the maximum within-pair desirability gaps were 1.635–2.097 for sentence-format forms and 1.598–2.399 for adjective-format forms, compared with 0.978 and 1.280 for the corresponding human optima. Conversely, when the human-selected forms were evaluated under the corresponding LLM-rating objectives, the maximum gaps were 2.000–3.000 for sentence-format forms and 2.467–3.267 for adjective-format forms. These results indicate that forms optimized under one rating source yielded less favorable objective values when evaluated under the other rating source, suggesting that rating-source-specific residual differences, in addition to optimization multiplicity, contributed to pair-level divergence.

We also confirmed that z-standardizing the LLM ratings did not change the selected item pairs in any of the six commercial-model conditions, and that linearly matching the LLM rating mean and standard deviation to the corresponding human ratings left the selected pairs unchanged in the two conditions carried forward to Stage III. Taken together, these results indicate that the low exact-pair overlap should be interpreted in light of both constrained optimization with many alternative optimal solutions and rating-source-specific residual differences, rather than as evidence that the LLM-based forms failed to satisfy the optimization criteria.

I Exploratory Open-Weight Model Analyses

To examine whether the patterns observed for the commercial API models were also observed for open-weight models accessed through OpenRouter, we conducted a post-review exploratory analysis using Meta Llama 3.3 70B Instruct and Mistral Large 3 2512. These analyses were not part of the preregistered primary analysis. They were intended as supplementary checks of whether the rating-level

and scale-construction patterns observed for the commercial API models also appeared for open-weight models. The analyses were limited to Stage I rating correspondence and Stage II scale construction; no Stage III human-participant validation was conducted for the open-weight-model-based forms.

I.1 Model and data-collection settings

For Llama, the model was Meta Llama 3.3 70B Instruct, using the OpenRouter model ID `meta-llama/llama-3.3-70b-instruct`. The provider was DeepInfra, with provider route `deepinfra/turbo`. The acquisition timestamp recorded in the prerun manifest corresponded to July 13, 2026. For Mistral, the model was Mistral Large 3 2512, using the OpenRouter model ID `mistralai/mistral-large-2512`. The provider was Mistral, with provider route `mistral`. The acquisition timestamp recorded in the prerun manifest corresponded to July 14, 2026. For both models, 15 requests were submitted per item, no generation parameters were explicitly set, and the acquisition scripts used the same item-level prompt template and seven-point response scale as the main Stage I rating task.

For both models, the analysis included 60 sentence items and 59 adjective items after excluding adjective Item 44 (*sicure*) because of the source-material discrepancy described in Appendix F.6. All 119 analysis-eligible items had 15 valid responses per item for each model. For Llama, 13 terminal invalid responses remained after the prespecified content-refit procedure, all associated with this pre-excluded item. These responses were not transformed or imputed, were not included in the analyses, and were not subjected to additional or replacement analyses. For Mistral, no terminal invalid responses were recorded.

I.2 Stage I rating correspondence and Stage II construction results

Stage I rating correspondence was summarized using Pearson's r, its 95% confidence interval, and the concordance correlation coefficient (CCC). All four open-weight model-by-format conditions exceeded the preregistered numerical benchmark of $r \geq .70$. In Stage II, the open-weight model ratings were entered into the same forced-choice scale-construction procedure used in the main analyses, with the same structural constraints and optimization procedure. The table reports item overlap, exact-pair overlap, and the standardized excess human-rated minimax gap. Sentence-format analyses used $N = 60$ items, and adjective-format analyses used $N = 59$ items after excluding Item 44. Each selected form consisted of 42 items and 21 item pairs. Table 13 summarizes the Stage I rating-correspondence results and the Stage II scale-construction outcomes for the two open-weight models.

The standardized excess human-rated minimax gap was defined as

$$G_{m,t} = \frac{M_H(P_{m,t}) - M_H(P_{H,t})}{\mathrm{SD}\left\{M_H(P_{b,t})\right\}_{b=1}^{100}},$$

where $P_{m,t}$ denotes the form constructed from the ratings of model m for format t, $P_{H,t}$ denotes the human-rating-based optimum for format t, $M_H(P)$ denotes the maximum within-pair desirability difference when form P is evaluated under human ratings, and $P_{b,t}$ denotes one of 100 feasible reference forms satisfying the same structural constraints. Smaller values indicate human-rated minimax performance closer to the human-based optimum.

For each format, the 100 reference forms were generated from the human ratings by repeatedly solving the original structural-constraint system with independently drawn continuous uniform weights on candidate pairs as the objective. Duplicate selected-pair sets were discarded until 100 unique feasible forms were retained. Thus, these forms provide a randomized feasible-reference distribution; they are not a uniform sample of all feasible forms.

For the three commercial API models, the standardized excess human-rated minimax gap ranged from 1.874 to 3.191 for sentence-format forms and from 1.199 to 4.220 for adjective-format forms. The Llama-based forms fell below the corresponding commercial-model ranges in both formats. The Mistral sentence-format form fell within the corresponding commercial-model range, whereas the Mistral adjective-format form exceeded it.

I.3 Interpretation and limitations

Overall, rating-level correspondence was high for both open-weight models, whereas Stage II outcomes were model- and format-dependent. Because these analyses did not include Stage III human-participant validation, they should be interpreted only as exploratory evidence about rating-level correspondence and propagation through

Model	Format	r [95% CI]	CCC	Item overlap	Exact-pair overlap	Std. excess gap
Llama 3.3 70B	Sentence	.946 [.911, .968]	.832	31/42	0/21	1.760
Llama 3.3 70B	Adjective	.978 [.963, .987]	.915	27/42	0/21	0.999
Mistral Large 3 2512	Sentence	.879 [.804, .926]	.796	33/42	0/21	2.635
Mistral Large 3 2512	Adjective	.918 [.865, .950]	.899	28/42	1/21	8.299

Table 13: Exploratory open-weight Stage I rating correspondence and Stage II construction results.

Stage II scale construction, not as evidence that open-weight-model-based forms preserve latent trait estimates.

J Results of the Exploratory Analyses in Stage III

As an exploratory analysis, Bland–Altman plots were inspected to examine potential systematic bias and dispersion patterns beyond correlation-based indices. Figure 5 presents Bland–Altman plots comparing latent trait estimates derived from human-based and LLM-based scales for sentence and adjective formats. Across both scale formats, the mean differences between the two estimates were close to zero, indicating little systematic bias.

Most observations fell within the limits of agreement, and the spread of differences was relatively consistent across the range of mean values. No clear pattern of increasing or decreasing differences was observed as a function of the magnitude of the latent trait estimates.

The limits of agreement were slightly wider for adjective items than for sentence items, suggesting somewhat greater variability in the differences for adjective-based scales. However, in both formats, the overall pattern suggests that differences between human- and LLM-based estimates were generally small, with no clear evidence of systematic bias.

J.1 Additional Stage III Agreement and Score-Precision Diagnostics

As exploratory analyses, we computed additional summaries of measurement-level agreement between the human-based and LLM-based forms. These analyses complemented the preregistered latent-trait correlation analyses and were not used as confirmatory decision criteria.

First, concordance correlation coefficients (CCCs) were computed to summarize absolute agreement in latent trait estimates. Across the Stage III comparisons, CCC values ranged from 0.657 to 0.844. Mean differences between LLM-based and human-based latent trait estimates ranged from -0.012 to 0.021, indicating little average directional bias.

Second, we examined posterior standard deviations of latent trait estimates as a diagnostic of score uncertainty. The mean posterior SD was 0.664 for the human-based forms and 0.687 for the LLM-based forms. These values should be interpreted as posterior uncertainty summaries, not as reliability coefficients. Conditional empirical expected a posteriori (EAP) reliability was also computed as an exploratory score-quality diagnostic, defined as the variance of posterior mean trait estimates divided by the sum of that variance and the mean posterior variance. Across HEXACO domains, reliability ranged from 0.531 to 0.702 for sentence-format human-based forms, from 0.485 to 0.692 for sentence-format LLM-based forms, from 0.381 to 0.701 for adjective-format human-based forms, and from 0.429 to 0.638 for adjective-format LLM-based forms. Thus, the posterior SD summaries complement, rather than replace, these reliability estimates.

Finally, we examined item-parameter stability for items shared between the human-based and LLM-based forms. Specifically, we compared the discrimination magnitude a_j^+ for shared items across forms as a descriptive parameter-stability check. The shared-item discrimination correlations were 0.905 across 33 shared sentence-format items and 0.806 across 30 shared adjective-format items. These analyses provide additional descriptive evidence about shared-item parameter stability, but they are not formal measurement-invariance tests.

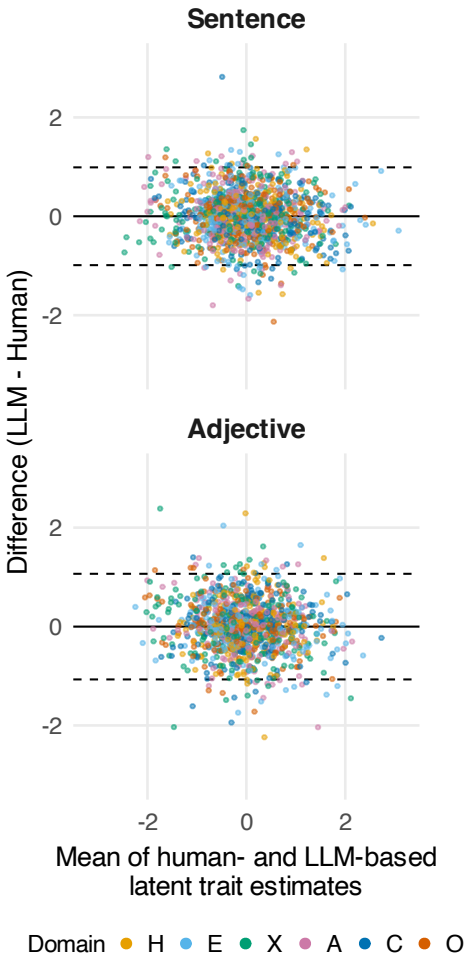


Figure 5: Bland–Altman plots of latent trait estimates from the human-based and LLM-based scales for the sentence (top) and adjective (bottom) formats. Each point represents one participant-by-domain comparison; colors indicate the six HEXACO domains (H, E, X, A, C, and O). The horizontal axis shows the mean of the human-based and LLM-based latent trait estimates, and the vertical axis shows their difference, computed as LLM-based minus human-based. The solid horizontal line indicates the mean difference within each format, and the dashed horizontal lines indicate the limits of agreement defined as the mean difference ± 1.96 SD.