Quantifying and Mitigating Socially Desirable Responding in LLMs: A Desirability-Matched Graded Forced-Choice Psychometric Study

Anonymous ACL submission

Abstract

Human self-report questionnaires are increasingly used in NLP to benchmark and audit large language models (LLMs), from persona consistency to safety and bias assessments. Yet these instruments presume honest responding; in evaluative contexts, LLMs can instead gravitate toward socially preferred answers—a form of socially desirable responding (SDR)—biasing questionnaire-derived scores and downstream conclusions. We propose a psychometric framework to quantify and mitigate SDR in questionnaire-based evaluation of LLMs. To quantify SDR, the same inventory is administered under HONEST versus FAKE-GOOD instructions, and SDR is computed as a direction-corrected standardized effect size from item response theory (IRT)-estimated latent scores. This enables comparisons across constructs and response formats, as well as against human instructed-faking benchmarks. For mitigation, we construct a graded forced-choice (GFC) Big Five inventory by selecting 30 cross-domain pairs from an item pool via constrained optimization to match desirability. Across nine instruction-tuned LLMs evaluated on synthetic personas with known target profiles, Likert-style questionnaires show consistently large SDR, whereas desirability-matched GFC substantially attenuates SDR while largely preserving the recovery of the intended persona profiles. These results highlight a model-dependent SDR–recovery trade-off and motivate SDR-aware reporting practices for questionnaire-based benchmarking and auditing of LLMs.[1]

1 Introduction

Large language models (LLMs) are increasingly evaluated not only for task performance but also for higher-level behavioral and psychological profiles. Benchmarks for cognitive capabilities and task performance in LLMs are now abundant. By contrast, despite the growing prominence of measuring and cultivating *non-cognitive* traits in human development (Heckman and Kautz, 2012; OECD, 2021), methods to measure and evaluate analogous *non-cognitive* attributes in LLMs—such as interpersonal style, value-laden tendencies, and stable response patterns—remain less standardized.

Psychometrics is the measurement science that develops and validates statistical models for inferring latent psychological traits from assessment responses. Since its early twentieth-century origins at the intersection of psychology and statistics (e.g., Spearman, 1904), the field has established modern measurement frameworks, most notably item response theory (IRT; van der Linden, 2018a).

Recent LLM studies adapt psychometric inventories, such as Big Five personality questionnaires, as a pragmatic, theory-grounded tool for non-cognitive evaluation, assessing stable trait-like signatures in LLM outputs (Serapio-García et al., 2023; Ye et al., 2025; Bhandari et al., 2025). Throughout, we interpret questionnaire scores as *behavioral* response tendencies under standardized prompts, rather than as evidence that LLMs possess human-like inner dispositions. These studies open the door to *AI psychometrics* (Pellert et al., 2024, 2025), but they also inherit some challenges from human assessment. A key challenge is *socially desirable responding* (SDR), in which respondents tend to select options they believe are socially preferred or likely to impress evaluators, rather than those that reflect their true traits (Paulhus, 2002). This behavior can distort self-report results and reduce their validity, as inflated scores obscure genuine individual characteristics.

Recent studies have indeed observed that LLMs also produce questionnaire responses with a strong prosocial or idealized bias, much like those of a human respondent (Bodroža et al., 2024; Salecha et al., 2024). This is unsurprising, given that most

[1] Data and code available at: https://osf.io/2e6ny/?view_only=678f4dbe220c4a1793eb3e159fcbc3f6

LLMs are fine-tuned to be helpful and avoid offending, which is a form of *alignment* with human preferences. Yet current LLM studies often report questionnaire-derived score profiles without quantifying the magnitude of SDR, without calibrating it to human benchmarks, and without deploying measurement formats designed to mitigate SDR. This raises the question of whether—and to what extent—questionnaire-derived proxies of personality constructs in LLM outputs are biased by social desirability, and how to design evaluations that yield more robust, better-isolated measurements.

A substantial body of literature in human psychometrics shows that forced-choice formats reduce faking compared to single-stimulus Likert questionnaires, especially in high-stakes contexts (Cao and Drasgow, 2019; Martínez and Salgado, 2021; Speer et al., 2023). Crucially, this advantage requires IRT-based scoring of comparative data to avoid ipsativity and to place respondents on a common latent scale needed for our SDR quantification and mitigation. (See Section 3.3 for more details; Brown and Maydeu-Olivares, 2011; Brown, 2016). This paper transfers these principles to psychometric profiling of LLM outputs under questionnaire-style prompts. Specifically, we treat SDR as a *measurable response distortion* that can be quantified through instruction-induced contrasts and mitigated through desirability-controlled comparative measurement. While design theory for desirability-matched comparative questionnaires is well developed, GFC Big Five instruments with explicit desirability matching are not commonly available. To advance psychometrically grounded evaluation and mitigation of socially desirable responding in questionnaire-based LLM assessments, our contributions are threefold.

Contributions.

1. **Quantifying socially desirable responding:** Inspired by SDR quantification in human assessment, we propose a psychometrically grounded SDR metric for LLMs—an instruction-induced effect size computed on IRT-based latent scores.
2. **A desirability-matched GFC Big Five inventory:** We build a GFC Big Five inventory by selecting desirability-matched cross-domain statement pairs via constrained optimization grounded in forced-choice psychometrics.
3. **Empirical validation and exploratory model differences:** Across nine instruction-tuned LLMs, we show that desirability-matched GFC attenuates SDR relative to Likert while retaining ground-truth recovery (vs. persona targets), and we characterize how the remaining SDR–recovery trade-off varies across models.

2 Related Work

Questionnaire-based elicitation of trait-relevant response tendencies in LLMs. A growing body of work treats LLMs as respondents to psychometric questionnaires (e.g., Big Five) to characterize behavior and probe prompt-level controllability. Early studies profiled foundation models and compared scores to human norms (Miotto et al., 2022). Subsequent work shows that profiles can be *elicited* or *steered* via prompting and can influence downstream text generation (Jiang et al., 2024), but contextual cues alone can shift perceived personality, highlighting instruction- and context-dependence (Caron and Srivastava, 2023). A complementary line of work examines measurement *reliability*: minor prompt or formatting perturbations can change responses substantially (Shu et al., 2024; Gupta et al., 2024), though some stability is reported under extensive resampling (Huang et al., 2024). Recent efforts propose LLM-oriented designs, including open-ended linguistic assessment (Zheng et al., 2025) and psychometrically validated multi-choice benchmarks (Lee et al., 2025).

Socially desirable responding in LLM survey settings. Response distortion in LLM questionnaires is closely related to *sycophancy*-tuned assistants mirroring a user's stated beliefs rather than answering independently. Perez et al. (2023) identify sycophancy via model-written evaluations, and Sharma et al. (2024) analyze its prevalence and likely drivers. Although usually studied in open-ended dialogue, the same "agree-with-the-user" incentive is conceptually adjacent to survey settings: when a model infers what response is expected or preferred, it may favor socially palatable options over its baseline tendencies. Direct evidence of socially desirable responding (SDR) in LLM assessment supports this. Across Big Five inventories, LLMs detect evaluation contexts and shift toward socially desirable profiles, paralleling human SDR (Salecha et al., 2024); broader AI psychometrics likewise finds interpretable yet systematically skewed profiles shaped by training data, safety tuning, and instruction framing (Pellert et al., 2024). However, existing SDR analyses mostly report aggregate score shifts, providing limited ability to dis-

entangle trait change from social desirability and other response-style effects at the item-parameter level, which hinders calibrated comparisons across models, instruments, and against human benchmarks.

Reducing response bias with comparative formats. A promising mitigation is to replace independent Likert ratings with *comparative* judgments. Li et al. (2025b) compare Likert and forced-choice Big Five questionnaires for LLMs and find forced-choice to be less sensitive to temperature and to better differentiate models; they further explore self-reflection prompting and LLM-as-a-judge scoring. Yet forced-choice formats can still admit desirability-driven selections unless the compared statements are carefully matched in desirability, and existing implementations often rely on heuristic scoring rather than probabilistic trait estimation.

Taken together, prior work leaves open (i) how to quantify SDR in a way that is comparable across traits and anchored to human faking behavior, and (ii) how to design and score forced-choice inventories that reduce SDR while retaining interpretable trait scores. Our work addresses these gaps with an IRT-calibrated SDR metric and a desirability-matched graded forced-choice inventory.

3 Method

3.1 Overview

Our method quantifies SDR as an instruction-induced shift in IRT-estimated scores. Concretely, we administer the same inventory under two instruction conditions: HONEST (answer as you really are) and FAKE-GOOD (answer to make the best possible impression). We fit an IRT model to place all responses on a common latent scale and define SDR as the resulting effect size (standardized difference) in trait scores between the two conditions (Figure 1). This SDR effect size becomes the shared target quantity for our analyses: we test whether GFC responding suppresses desirability-driven distortion relative to standard Likert responding, and we conduct exploratory comparisons of SDR magnitudes across model families.

To enable SDR mitigation, we first construct a desirability-matched Big Five GFC scale by estimating item desirability and pairing cross-trait statements with closely matched desirability. This prevents models from uniformly selecting the "socially best" option and allows IRT scoring to recover interpretable trait scores under both HONEST and FAKE-GOOD conditions.

3.2 Item pool, desirability estimation, and inventory construction

We begin from Goldberg's public-domain IPIP Big-Five factor-marker inventory (Goldberg, 1999), which contains 100 statements covering five domains. Following our data-collection protocol, we exclude two voting-related statements and retain $J = 98$ items. Each statement j is annotated with a Big Five domain label $f(j) \in \{A, C, E, N, O\}$, which represents the domains of agreeableness, conscientiousness, extraversion, neuroticism, and openness to experience, and a keying direction $g_j \in \{-1, +1\}$. $g_j = +1$ indicates that the statement is positively keyed (endorsing the statement indicates a higher trait level), and $g_j = -1$ means it is negatively keyed.

A key ingredient for desirability matching is an item-level social desirability score s_j. We estimate s_j by prompting two strong LLMs (GPT-5 and Gemini 2.5 Pro) to act as "desirability raters" and judge how socially desirable the characteristic described by each statement is. In this process, to ensure comparability with established human desirability norms, we use the same instructional framing and response anchors that are standard in human social desirability rating studies, asking how desirable each trait/characteristic is for an adult person on a 9-point scale (Sankis et al., 1999; Coker et al., 2002). Moreover, because LLM-based desirability ratings could deviate from human norms, we validate convergent agreement against human social desirability norms from Britz et al. (2023) (Appendix B). These steps provide the desirability inputs needed for constructing a matched GFC inventory.

Constructing desirability-matched graded forced-choice pairs. We construct a graded forced-choice inventory with $P = 30$ paired-statement blocks (60 unique statements) by selecting *cross-domain* item pairs whose desirability scores s_j are as closely matched as possible. Pair selection is posed as a constrained two-stage mixed-integer optimization that (i) minimizes the worst within-pair desirability gap and then (ii) among minimax-optimal solutions, minimizes overall mismatch while balancing trait coverage, domain-pair counts, and keying composition (Sun et al., 2024; Li et al., 2025a). Appendix C provides

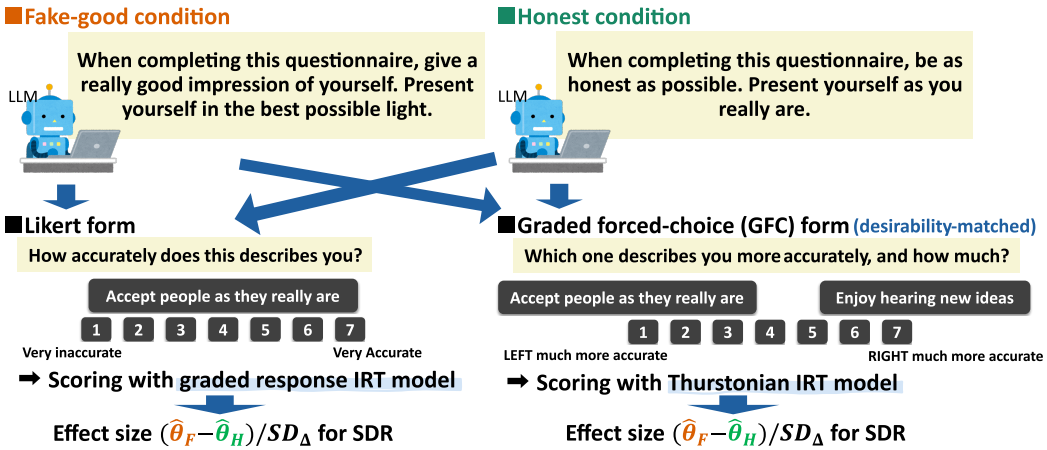

Figure 1: Proposed framework to quantify and compare the socially desirable responding (SDR) of LLMs.

the full formulation and the resulting inventory; the final set achieves a maximum within-block desirability gap of 0.18 on the 1–9 scale (mean 0.03).

Response formats. We administer the same 60 selected statements in two formats:

- LIKERT (SINGLE-STIMULUS): each statement is rated on a 7-point Likert scale (1 = very inaccurate, 7 = very accurate).
- GFC (PAIRWISE GRADED COMPARISON): statements are presented in 30 desirability-matched pairs; respondents provide a graded comparative judgment on a 7-point bipolar scale (1 = LEFT much more accurate, 7 = RIGHT much more accurate; Zhang et al., 2024).

3.3 Psychometric scoring models

Why IRT scoring? To measure the trait-like profiles of LLMs based on raw responses, we use the IRT framework. The IRT offers a measurement backbone that is both theoretically grounded and practically useful, which is why it is adopted in major educational and psychological measurement programs such as OECD PISA (OECD, 2025), NIH PROMIS (PROMIS, 2013; Rose et al., 2014), and the SAT (College Board, 2024).

IRT provides a principled *measurement model* that maps observed responses to IRT-estimated latent scores via interpretable item parameters, rather than treating raw observed scores as direct measurements of the construct of interest. It offers a broad set of benefits for measurement—from principled test scoring to uncertainty-aware latent trait estimation and scale linking (van der Linden, 2018b). In our setting, one of the most important advantages is pragmatic and methodological: IRT provides a single, coherent measurement backbone that we can apply consistently across two different response formats, Likert ratings and GFC. This common backbone is essential for our goal of comparing socially desirable responding (SDR) across instruction conditions and LLM families without conflating true response distortion with format-specific scoring artifacts.

Notably, forced-choice responses cannot be treated as ordinary observed scores. Naively scoring comparative choices, such as counting how often a statement or a choice option is selected within each block, produces *ipsative*, constant-sum scores, which do not support valid between-respondent comparisons and can induce spurious dependencies among trait estimates (Brown and Maydeu-Olivares, 2013). Psychometrics addresses this limitation by extending IRT models from Likert-type data to forced-choice designs, enabling the recovery of normative latent traits from comparative judgments. In particular, Thurstonian IRT provides a direct solution to ipsativity by positing latent utilities for each statement and modeling observed choices as differences between these utilities (Brown and Maydeu-Olivares, 2011; Brown, 2016; Bürkner, 2022). This yields latent trait estimates on a normative scale that are comparable across respondents and conditions—exactly what is required to quantify SDR as a systematic shift in latent traits rather than a byproduct of how responses are elicited.

Likert: five-dimensional graded response model (GRM). We score both formats with appropriate multidimensional IRT models fitted in Stan (Car-

penter et al., 2017). Let i index a response unit, which corresponds to a persona in our experiment (see Section 4.2). Let $\boldsymbol{\theta}_i \in \mathbb{R}^5$ denote its latent Big Five trait score vector. Let $\boldsymbol{q}_j \in \{0,1\}^5$ be the one-hot vector indicating the measured trait for item j. Let $Y_{ij} \in \{1,\dots,7\}$ be unit i's Likert response to item j. We use a multidimensional graded response model (Samejima, 1969), which is a standard IRT model for observed ordinal (Likert-type) item responses. Define a signed discrimination $a_j = g_j a_j^+$ with $a_j^+ > 0$ and keying sign g_j. The linear predictor is

$$\eta_{ij} = a_j \boldsymbol{q}_j^\top \boldsymbol{\theta}_i. \tag{1}$$

With item-specific ordered thresholds $\kappa_j = (\kappa_{j1},\dots,\kappa_{j6})$ satisfying $\kappa_{j1} < \kappa_{j2} < \cdots < \kappa_{j6}$, the GRM can be written as

$$\Pr(Y_{ij} \geq k \mid \boldsymbol{\theta}_i) = \text{logit}^{-1}(\eta_{ij} - \kappa_{j,k-1}), \quad k = 2,\dots,7. \tag{2}$$

GFC: logistic ordinal Thurstonian IRT. For each pair $p \in \{1,\dots,30\}$ with the left statement L_p and the right statement R_p, let $Y_{ip} \in \{1,\dots,7\}$ denote the graded comparative response. Following the literature in the ordinal Thurstonian IRT model (Brown, 2016; Brown and Maydeu-Olivares, 2018; Bürkner, 2022; Okada and Bunji, 2021), we define each statement's latent utility

$$\mu_{ij} = a_j \boldsymbol{q}_j^\top \boldsymbol{\theta}_i, \tag{3}$$

and the comparative signal as the scaled right-minus-left difference

$$\eta_{ip} = \frac{\mu_{i,R_p} - \mu_{i,L_p}}{\sqrt{2}}. \tag{4}$$

We then model the graded comparison with pair-specific ordered thresholds $\kappa_p = (\kappa_{p1},\dots,\kappa_{p6})$ satisfying $\kappa_{p1} < \kappa_{p2} < \cdots < \kappa_{p6}$:

$$\Pr(Y_{ip} \geq k \mid \boldsymbol{\theta}_i) = \text{logit}^{-1}(\eta_{ip} - \kappa_{p,k-1}), \quad k = 2,\dots,7. \tag{5}$$

so that larger η_{ip} implies stronger endorsement of the RIGHT statement.

3.4 Quantifying socially desirable responding

Our SDR metric operationalizes instructed faking as an instruction-induced shift in latent trait estimates. For each LLM[2], we obtain latent trait estimates under HONEST and FAKE-GOOD instructions for each response format $f \in \{\text{LIKERT}, \text{GFC}\}$. For each Big Five trait $t \in \{A, C, E, N, O\}$, let $\hat{\theta}_{i,t,\text{HONEST}}$ and $\hat{\theta}_{i,t,\text{FAKE}}$ denote the posterior-mean latent trait estimates.

We define the raw within-persona shift for trait t as

$$\Delta_{i,t} = \hat{\theta}_{i,t,\text{FAKE}} - \hat{\theta}_{i,t,\text{HONEST}}. \tag{6}$$

We compute the paired standardized effect size across personas (Cohen's d_z for dependent means) (Cohen, 1988; Lakens, 2013):

$$d_{z,t} = \frac{\overline{\Delta}_{\cdot,t}}{SD_{\Delta_{i,t}}}, \tag{7}$$

where $\overline{\Delta}_{\cdot,t}$ is the mean shift across personas and $SD_{\Delta_{i,t}}$ is its standard deviation.

In Big Five measurement, higher Agreeableness, Conscientiousness, Extraversion, and Openness are generally socially desirable, whereas lower Neuroticism is desirable. To make interpretation uniform across traits, we direction-correct signs so that positive values always indicate movement toward social desirability. We define a desirability-direction multiplier g_t as $g_t = +1$ for $t \in \{A, C, E, O\}$ and $g_t = -1$ for $t = N$. Thus, our direction-corrected effect size is $\tilde{d}_{z,t} = g_t\, d_{z,t}$.

3.5 Ground-truth recovery metric

Because LLM "personas" have known ground-truth trait profiles by construction, we evaluate ground-truth recovery by correlating estimated traits with persona specifications. For each persona i, we have a target Big Five vector $\boldsymbol{z}_i = (z_{A,i}, z_{C,i}, z_{E,i}, z_{N,i}, z_{O,i})$ used to generate persona descriptions (Section 4.1). For response format f and instruction condition $c \in \{\text{HONEST}, \text{FAKE-GOOD}\}$, we compute the Pearson correlation between the estimated trait vector $\hat{\boldsymbol{\theta}}_{i,c}$ and the target $\boldsymbol{z}_i$ across personas, and report trait-wise and averaged correlations. This captures whether measurement preserves rank-order differences in intended personality profiles.

4 Experimental Setup

4.1 Personas and ground-truth traits

To mimic an LLM adopting diverse user personas and to evaluate how well the IRT models recovered trait levels, we generated a shared set of 50 synthetic personas with known Big Five trait values and used it for all LLMs. We sample $\boldsymbol{z}_i$ from a multivariate standard normal distribution with

[2]For readability, we suppress the LLM index l in what follows; all quantities are understood to be computed for a fixed LLM unless l is shown explicitly.

mean zero and an empirical human Big Five correlation structure adopted from van der Linden et al. (2010) so that the sampled personas resemble realistic joint distributions of real-world Big Five profiles. (Appendix E). We then map z_i to trait adjectives by selecting descriptors from a curated trait lexicon, ensuring coverage across the five domains.

For constructing natural language descriptions that are more amenable to processing by LLMs, we transformed each trait score into persona descriptions based on the 52 markers from Serapio-García et al. (2023). Specifically, for each trait, we produced one sentence listing adjectives/phrases consistent with the assigned stanine scores of the traits with intensity terms such as "extremely", "very", or "a bit" depending on the stanine score of that trait. Each persona description always covered all five traits and was framed by an explicit role instruction ("YOU ARE THE RESPONDENT") and a final directive to answer as that person would. In this way, we achieved both the assignment of ground-truth Big Five scores and the imitation of adopted personas that are realistic and amenable to LLMs.

4.2 Experimental design, models, and prompting

We evaluate nine instruction-tuned LLMs from three providers: OpenAI (GPT-5, GPT-5 mini, GPT-5 nano), Google (Gemini 2.5 Pro, Gemini 2.5 Flash, Gemini 2.5 Flash-Lite), and Anthropic (Claude Opus 4.5, Claude Sonnet 4.5, Claude Haiku 4.5). For each persona and LLM, we administer both response formats under a fully crossed within-respondent design with two factors: (i) instruction condition (HONEST vs. FAKE-GOOD) and (ii) response format (LIKERT vs. GFC). Thus each persona yields four questionnaire response sets per model.

Instruction manipulation and prompting. To elicit SDR, we administer each questionnaire under two instruction conditions. In the HONEST condition, the prompt instructed the persona to answer as honestly as possible. In the FAKE-GOOD condition, the prompt instructed the persona to answer in order to give a really good impression and to present themselves in the best possible light. These instructions were adopted from Furnham (1997), which introduced a faking-instructions paradigm that has since been widely used to study intentional response distortion.

Prompts present the persona description, followed by the questionnaire items or pairs, and require the model to output only the requested numeric responses. Full prompts are provided in Appendix F.

4.3 Data collection and IRT estimation

We collect complete response vectors for each (LLM, persona, format, condition) combination. Within each persona, item/pair order is randomized once and held fixed across instruction conditions to support within-persona comparisons. All LLMs are queried via their official APIs in batch mode.

We fit the GRM and ordinal Thurstonian IRT models in Stan (Carpenter et al., 2017) with weakly informative priors and standard identifiability constraints (Appendix D).

5 Results

5.1 Profile distortion under fake-good instructions

Figure 2 visualizes the fake-good distortion observed in our experiments for GPT-5 and Gemini 2.5 Pro, demonstrating how the response format shapes the distortion. Under the single-stimulus Likert format, fake-good instructions pull the inferred Big Five profile toward a stereotypically "good person" pattern: higher A, C, E, and O, and lower N. The fake-good (orange) polygon expands on socially valued traits and contracts on Neuroticism relative to the honest (green) polygon. By contrast, the GFC panels show substantially smaller separations between honest and fake-good profiles for the same models. When an LLM must make graded comparative judgments between (approximately) equally desirable statements, it becomes difficult to uniformly endorse all "good" content. As a result, the honest and fake-good profiles in GFC are much closer than their Likert counterparts.

5.2 SDR magnitude under Likert vs. desirability-matched GFC

To quantify the magnitude of SDR across models, Figure 3 summarizes the effect sizes on the latent Big Five trait estimates ($\hat{\theta}$) for each response format. The heatmap reports direction-corrected Cohen's d_z effect sizes (fake-good minus honest; Cohen, 1988; Lakens, 2013), which is the paired-samples standardized mean difference, oriented so that positive values always indicate a shift toward the socially desirable direction (higher A, C, E, O, and lower N).

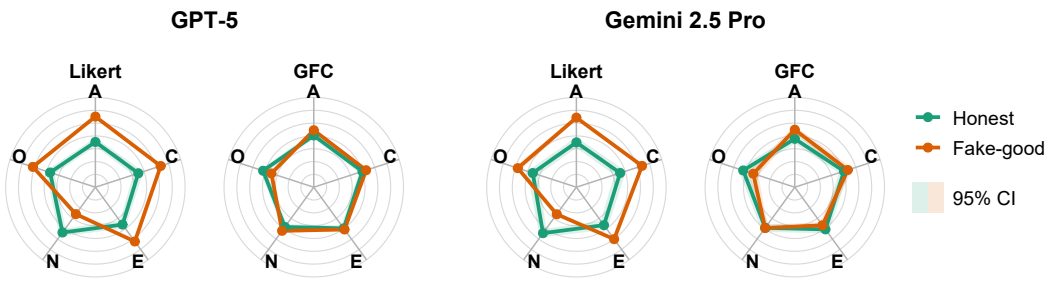

Figure 2: Mean ±95% CI Big Five profiles for two representative high-capacity LLMs (GPT-5 and Gemini 2.5 Pro) under HONEST vs. FAKE-GOOD instructions. For each model, trait estimates are shown separately for Likert responses (left) and graded forced-choice (GFC) responses (right).

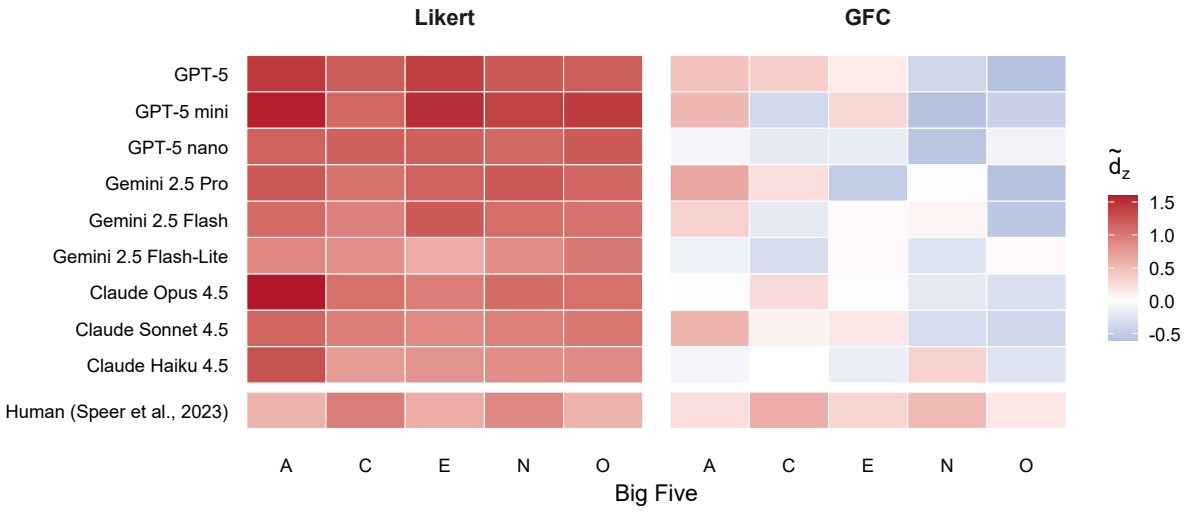

Figure 3: SDR effects on Big Five trait estimates (direction-corrected Cohen's $\tilde{d}_z$ computed on latent $\hat{\theta}$). Positive values indicate shifts toward the socially desirable direction (higher A, C, E, and O; lower N). The Likert panel shows consistently large, positive SDR across models and traits, whereas the GFC panel shows substantially attenuated (often near-zero) effects.

The Likert panel of Figure 3 shows a strikingly consistent pattern: for every evaluated model, fake-good instructions produce positive shifts on all five traits once aligned to desirability. This indicates that Likert responses allow LLMs to substantially reshape their inferred personality profiles when prompted to "look good." In other words, on Likert items the fake-good manipulation behaves like a global response-style shift toward desirability rather than a subtle or trait-specific perturbation. These magnitudes are comparable to human instructed-faking effects summarized in meta-analyses (Speer et al., 2023).

In contrast, desirability-matched GFC substantially attenuates SDR across most models. The reduction is strongest for models that show the largest Likert SDR, suggesting that desirability matching and comparative judgment jointly constrain impression management. However, some models retain non-negligible SDR even under GFC, indicating that comparative formats do not fully eliminate desirability-driven response distortion in LLMs.

5.3 Recovery of persona ground truth

SDR attenuation alone is not sufficient because a response format could trivially yield near-zero SDR by suppressing trait-relevant variance, leaving the resulting trait estimates uninterpretable. Because our personas have ground-truth target trait vectors by construction, we therefore check whether estimated traits preserve this signal via correlations with the persona ground truth.

Figure 4 summarizes the resulting SDR–recovery relationship. Under Likert, models of-

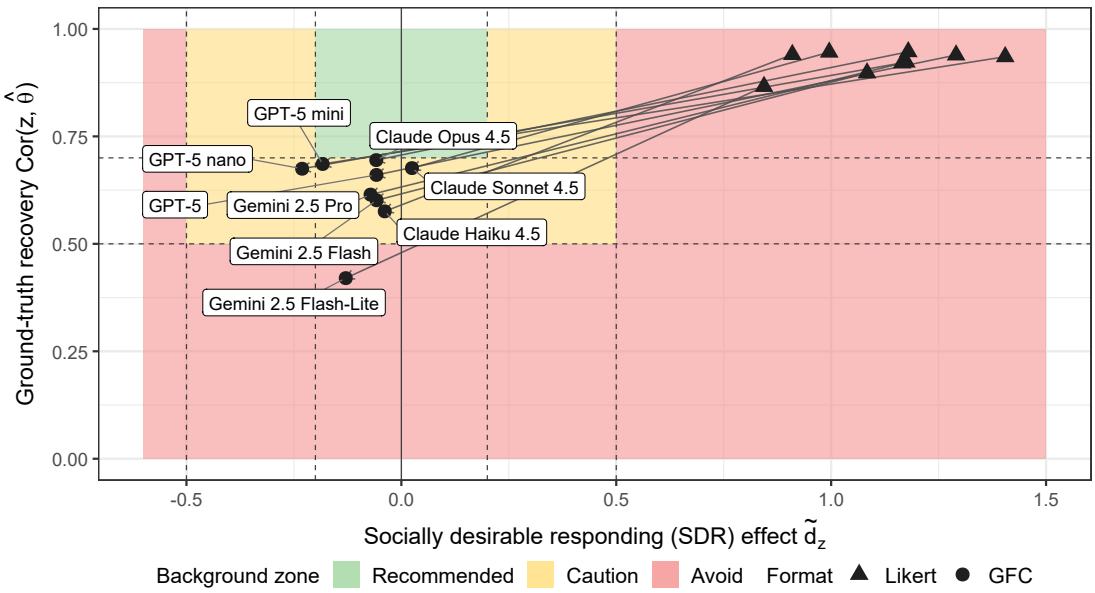

Figure 4: SDR–recovery trade-off across response formats. For each LLM and format, we plot summary SDR shift (direction-corrected Cohen's $\tilde{d}_z$) against ground-truth recovery (Pearson r between true persona scores z and IRT-estimated $\hat{\theta}$) under honest responding, both aggregated across traits. Grey lines connect formats within a model. Background colors indicate practical interpretation zones: for SDR, $|\tilde{d}_z| \leq 0.2$ (Cohen's (1988) "small" effect) is considered a practically negligible shift (recommended), $0.2 < |\tilde{d}_z| \leq 0.5$ (up to "medium") as caution but potentially acceptable, and $|\tilde{d}_z| > 0.5$ as avoid; these cutoffs are adopted from the equivalence-testing literature discussing equivalence bounds (smallest effect size of interest, SESOI) for assessing practically negligible differences (Lakens et al., 2018). For recovery, following the convergent-validity literature, we interpret $r \geq 0.70$ as strong (recommended), $0.50 \leq r < 0.70$ as acceptable/moderate, and $r < 0.50$ as insufficient (Abma et al., 2016).

ten achieve higher ground-truth recovery, but their SDR shifts are consistently large, placing most points in the red (avoid) region—meaning that in self-presentational contexts, the inferred trait profiles can be substantially distorted by desirability-driven responding. The dominant effect of our proposed desirability-matched GFC is therefore the large SDR reduction: for every model, switching from Likert to GFC produces a pronounced leftward shift, typically moving into the yellow/green (negligible-to-moderate SDR) region. This mitigation comes with a cost in ground-truth recovery, but the drop is generally limited: most GFC points remain in the acceptable-to-strong convergent-validity bands ($r \geq 0.50$), indicating that GFC suppresses desirability-driven distortion while largely preserving trait signal.

Although the general pattern is consistent, the degree of trade-off also varies by model family. These patterns imply that alignment and response policies interact with measurement format in model-specific ways. We qualitatively examine whether SDR magnitude and SDR–recovery trade-offs track coarse capability tiers within providers. While higher-tier models sometimes show lower SDR under GFC, the relationship is not uniform: some smaller models show minimal additional SDR beyond baseline, while some larger models retain measurable SDR under both formats. These findings motivate future work on linking SDR to specific training objectives and safety policies.

6 Conclusion

We proposed a psychometric framework to quantify and mitigate SDR in questionnaire-based profiling of LLM outputs. Building on SDR frameworks used in human psychometrics, we quantify SDR as a direction-corrected instruction-induced effect size on IRT-based latent traits. To reduce SDR, we develop a desirability-matched graded forced-choice Big Five inventory scored with ordinal Thurstonian IRT. Across nine instruction-tuned LLMs, Likert responses exhibit large SDR, whereas desirability-matched GFC substantially reduces SDR while largely preserving ground-truth recovery. Together, these findings motivate SDR-aware evaluation protocols that report both trait estimates and distortion metrics, enabling more reliable behavioral auditing across models and settings.

7 Limitations

A first set of limitations concerns how our empirical contrasts should be interpreted. In Figure 3, the human reference row is taken from a meta-analysis (Speer et al., 2023) that aggregates heterogeneous instruments, scoring approaches, and study contexts; as a result, the reported human effects do not necessarily reflect a single, desirability-matched GFC design scored with the same Thurstonian IRT model used in our study, and desirability is not always explicitly controlled in the primary studies. By contrast, our LLM results come from a single, tightly controlled design using the same statement pool, a desirability-matched GFC construction, and a shared IRT metric. Consequently, qualitative differences between humans and LLMs should be interpreted cautiously: they may reflect methodological heterogeneity in addition to intrinsic human–LLM differences. In this respect, our LLM experiments arguably provide a more controlled benchmark, but a direct, design-matched human–LLM study is needed to make stronger comparative claims.

In addition, we matched the *number of statements* across formats, but this implies that Likert collects twice as many responses as GFC (60 ratings vs. 30 pair judgments), potentially disadvantaging GFC in our ground-truth recovery comparison. Future work should therefore (i) administer the same desirability-matched inventory to humans and LLMs and score both with the same Thurstonian IRT model, and (ii) compare formats under matched response counts or matched test information.

A second set of limitations concerns the scope of the instrument and the nature of our ground truth. More broadly, our study focuses on a specific personality instrument (IPIP Big-Five factor-marker inventory; Goldberg, 1999). Although this inventory and the Big Five traits measured by it are considered representative of non-cognitive traits, other inventories may yield different SDR patterns. Our personas provide a controlled ground truth for recovery evaluation, but they are synthetic and may not reflect the full richness of human trait expression. Finally, we treat each persona–instruction instance as a separate response unit when fitting IRT; future work could explicitly model within-persona dependence and condition effects in a hierarchical framework.

8 Ethical Considerations

This work studies SDR in questionnaire-style profiling of LLMs and proposes a desirability-matched GFC inventory to reduce impression-management effects. The intended use is to improve the validity of behavioral evaluation and to support auditing settings (e.g., fairness/safety monitoring) where survey-style prompts are used and SDR can mask model differences.

At the same time, the proposed construction principles may be dual-use: they could be used to design questionnaires or prompts that elicit less detectable self-presentation, potentially enabling evaluation evasion or more persuasive synthetic personas. We therefore emphasize research and auditing use, report SDR–recovery trade-offs to make failure modes explicit, and encourage practitioners to avoid deploying personality-style scores for decision-making about people.

Our desirability matching relies on normative desirability judgments, which can be culturally and context dependent and may reflect rater-specific biases. Applications to new languages, cultures, or domains should re-validate desirability ratings and consider stakeholder-specific norms. Our experiments use synthetic personas and public-domain questionnaire items and do not involve personal data; any extension to human assessment should follow established ethics review and privacy protections.

References

Inger L. Abma, Maroeska Rovers, and Philip J. van der Wees. 2016. Appraising convergent validity of patient-reported outcome measures in systematic reviews: constructing hypotheses and interpreting outcomes. *BMC Research Notes*, 9:226.

Pranav Bhandari, Usman Naseem, Amitava Datta, Nicolas Fay, and Mehwish Nasim. 2025. Evaluating personality traits in large language models: Insights from psychological questionnaires. In *Companion Proceedings of the ACM Web Conference 2025*, WWW Companion '25, pages 868–872, New York, NY, USA. Association for Computing Machinery.

Bojana Bodroža, Bojana M. Dinić, and Ljubiša Bojić. 2024. Personality testing of large language models: limited temporal stability, but highlighted prosociality. *Royal Society Open Science*, 11(10):240180.

Sara Britz, Jessica Heintz, 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. Published online 12 Aug 2022.

Anna Brown. 2016. Item response models for forced-choice questionnaires: A common framework. *Psychometrika*, 81(1):135–160.

Anna Brown and Alberto Maydeu-Olivares. 2011. Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, 71(3):460–502.

Anna Brown and Alberto Maydeu-Olivares. 2013. How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods*, 18(1):36–52.

Anna Brown and Alberto Maydeu-Olivares. 2018. Ordinal factor analysis of graded-preference questionnaire data. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(4):516–529.

Paul-Christian Bürkner. 2022. On the information obtainable from comparative judgments. *Psychometrika*, 87(4):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.

Graham Caron and Shashank Srivastava. 2023. Manipulating the perceived personality traits of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2370–2386, Singapore. Association for Computational Linguistics.

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.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2 edition. Lawrence Erlbaum Associates.

Linda Anne Coker, Douglas B. Samuel, and Thomas A. Widiger. 2002. Maladaptive personality functioning within the big five and the five-factor model. *Journal of Personality Disorders*, 16(5):385–401.

College Board. 2024. Assessment framework for the digital SAT suite. Version 3.01 (August 2024).

Adrian F. Furnham. 1997. Knowing and faking one's five-factor personality score. *Journal of Personality Assessment*, 69(1):229–243.

Lewis R. Goldberg. 1999. A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. J. Deary, F. De Fruyt, and F. Ostendorf, editors, *Personality Psychology in Europe*, volume 7, pages 7–28. Tilburg University Press, Tilburg, The Netherlands.

Akshat Gupta, Xiaoyang Song, and Gopala Anumanchipalli. 2024. Self-assessment tests are unreliable measures of LLM personality. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 301–314, Miami, Florida, US. Association for Computational Linguistics.

James J. Heckman and Tim Kautz. 2012. Hard evidence on soft skills. *Labour Economics*, 19(4):451–464.

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.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.

Daniël Lakens. 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and anovas. *Frontiers in Psychology*, 4:863.

Daniël Lakens, Anne M. Scheel, and Peder M. Isager. 2018. Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2):259–269.

Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, Jinyoung Yeo, and Youngjae Yu. 2025. Do LLMs have distinct and consistent personality? TRAIT: Personality testset designed for LLMs with psychometrics. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8397–8437, Albuquerque, New Mexico. Association for Computational Linguistics.

Mengtong Li, Bo Zhang, Lingyue Li, Tianjun Sun, and Anna Brown. 2025a. Mix-keying or desirability-matching in the construction of forced-choice measures? an empirical investigation and practical recommendations. *Organizational Research Methods*, 28(2):296–329.

Xiaoyu Li, Haoran Shi, Zengyi Yu, Yukun Tu, and Chanjin Zheng. 2025b. Decoding LLM personality measurement: Forced-choice vs. Likert. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9234–9247, Vienna, Austria. 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.

Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. 2022. Who is GPT-3? an exploration of personality, values and demographics. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 218–227, Abu Dhabi, UAE. Association for Computational Linguistics.

OECD. 2021. *Beyond Academic Learning: First Results from the Survey of Social and Emotional Skills*. OECD Publishing, Paris.

OECD. 2025. Pisa 2022 technical report (revised edition). Technical report, OECD Publishing, Paris.

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

Delroy L. Paulhus. 2002. Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson, and D. E. Wiley, editors, *The role of constructs in psychological and educational measurement*, pages 49–69. Erlbaum, Mahwah, NJ.

Max Pellert, Clemens M. Lechner, Indira Sen, and Markus Strohmaier. 2025. Neural network embeddings recover value dimensions from psychometric survey items on par with human data. *arXiv preprint*.

Max Pellert, Clemens M. Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2024. AI psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 19(5):808–826.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, and 44 others. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada. Association for Computational Linguistics.

PROMIS. 2013. Promis® instrument development and validation scientific standards. Version 2.0 (revised May 2013). Patient-Reported Outcomes Measurement Information System (PROMIS), funded by NIH.

Matthias Rose, Jakob B. Bjorner, Barbara Gandek, Bonnie Bruce, James F. Fries, and John E. Ware. 2014. The PROMIS physical function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. *Journal of Clinical Epidemiology*, 67(5):516–526.

Aadesh Salecha, Molly E. Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H. Ungar, and Johannes C. Eichstaedt. 2024. Large language models display human-like social desirability biases in big five personality surveys. *PNAS Nexus*, 3(12):pgae533.

Fumiko Samejima. 1969. Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, 34(Suppl 1):1–97. Psychometrika Monograph Supplement, No. 17.

Lizabeth M. Sankis, Elizabeth M. Corbitt, and Thomas A. Widiger. 1999. Gender bias in the English language? *Journal of Personality and Social Psychology*, 77(6):1289–1295.

Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja J. Matarić. 2023. Personality traits in large language models. *arXiv preprint*, arXiv:2307.00184.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*. ArXiv:2310.13548.

Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, and David Jurgens. 2024. You don't need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5263–5281, Mexico City, Mexico. Association for Computational Linguistics.

Charles Spearman. 1904. "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology*, 15(2):201–292.

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.

Luning Sun, Zijie Qin, Shan Wang, Xuetao Tian, and Fang Luo. 2024. Contributions to constructing forced-choice questionnaires using the thurstonian IRT model. *Multivariate Behavioral Research*, 59(2):229–250.

Dimitri van der Linden, Jan te Nijenhuis, and Arnold B. Bakker. 2010. The general factor of personality: A meta-analysis of big five intercorrelations and a criterion-related validity study. *Journal of Research in Personality*, 44(3):315–327.

Wim J. van der Linden, editor. 2018a. *Handbook of Item Response Theory: Three-Volume Set*. Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences. Chapman and Hall/CRC.

Wim J. van der Linden, editor. 2018b. *Handbook of Item Response Theory: Volume 3: Applications*. Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences. Chapman and Hall/CRC, Boca Raton, FL.

Haoran Ye, Jing Jin, Yuhang Xie, Xin Zhang, and Guojie Song. 2025. Large language model psychometrics: A systematic review of evaluation, validation, and enhancement. *arXiv preprint*.

Bo Zhang, Jing Luo, and Jian Li. 2024. Moving beyond likert and traditional forced-choice scales: A comprehensive investigation of the graded forced-choice format. *Multivariate Behavioral Research*, 59(3):434–460.

Jingyao Zheng, Xian Wang, Simo Hosio, Xiaoxian Xu, and Lik-Hang Lee. 2025. LMLPA: Language model linguistic personality assessment. *Computational Linguistics*, 51:599–640.

A Estimating social desirability scores for IPIP items

Overview We estimate item-level social desirability scores s_j for the IPIP item pool by prompting two rater LLMs (GPT-5 and Gemini 2.5 Pro) to rate how socially desirable each statement is for an adult person. The instruction framing and the 1–9 desirability anchors follow standard human social desirability rating protocols used in lexical and evaluative-norm studies (Sankis et al. 1999; Coker et al. 2002).

Prompt and Response Format Each rating query presented the model with a block of 25 personality statements. We adopt the same normative instruction wording and desirability anchors as in human ratings (Sankis et al., 1999; Coker et al., 2002), and add only the strict "return integers only" constraint to support automatic parsing. The general instruction (shared across blocks and rater models) read:

> The following statements are characteristics of people. Indicate on a scale from 1 to 9 **how desirable** you think each trait or characteristic is for an adult person: 1 = Very undesirable, 3 = Undesirable, 5 = Neutral, 7 = Desirable, 9 = Very desirable. Use any number from 1 through 9 (i.e., 1, 2, 3, 4, 5, 6, 7, 8, or 9) that best indicates your opinion regarding how desirable that trait is.
>
> Please return EXACTLY 25 integers separated by single spaces, in the SAME ORDER as the statements.
>
> Do not include any other text.

Below this instruction, the 25 items in the current block were listed line by line in the form `Statement: [sentence]` and visually delimited from the instruction text by marker lines ("++++"). Thus, the desired output for each query was a single line consisting of 25 integers in $\{1, \ldots, 9\}$, each representing the judged desirability of the corresponding statement.

Randomization and Batching For each rater model $l \in \{\text{GPT-5}, \text{Gemini 2.5 Pro}\}$ and each replication $r \in \{1, \ldots, 30\}$, we performed the following steps:

1. We drew a random permutation of the J statements without replacement.
2. The permuted sequence was partitioned into four consecutive blocks of 25 items each (blocks $b = 1, \ldots, 4$).
3. For each block b, we constructed a prompt as described above and submitted it to the rater model via the provider API using batch processing (OpenAI via `ellmer::batch_chat_text`; Gemini via the Gemini Batch API over REST).

Each prompt was issued as a single-turn request (no chat history). Across 30 replications and 4 blocks per replication, each rater model therefore evaluated $30 \times 4 = 120$ prompts and responded $120 \times 25 = 3000$ desirability ratings.

Quality Control and Refitting of Non-conforming Outputs Raw API responses were parsed and quality-checked to ensure they could be unambiguously aligned to the 25 statements in the corresponding prompt. In our implementation, we normalized each response by removing line breaks, commas, and whitespace, and then verified that the resulting string consisted of exactly 25 single-digit ratings and contained only digits in $\{1, \ldots, 9\}$. Responses failing this check were flagged as non-conforming.

For flagged prompts, we re-issued the query once using the same item block but with an augmented prefix reminding the model that (i) this is a psychometric rating task, (ii) some items may mention sensitive topics, and (iii) it must return only the requested integers in the requested format. At this step, there were no responses that failed the format check.

Aggregation to Item-level Desirability Scores Because the order of statements was randomized on every replication, we stored the ordered list of item identifiers included in each 25-item prompt

and aligned the returned ratings back to item IDs by position (1–25). Let x_{jlr} denote the rating assigned to item $j \in \{1, \ldots, 100\}$ by rater model l on replication r. We computed the final desirability score for item j as the mean of all available ratings,

$$s_j = \frac{1}{60} \sum_l \sum_r x_{jlr},$$

We used the mean s_j as the desirability input to the optimization procedure described in Section 3.2.

Agreement of observed desirability ratings. To check the consistency between ratings, treating items as targets and replications as raters, we computed a two-way random-effects, absolute-agreement intraclass correlation coefficient (ICC; ICC(A,1) for a single replication and ICC(A,30) for the mean across 30 replications). Within each model, agreement was extremely high (Gemini 2.5 Pro: ICC(A,1)=.980; GPT-5: ICC(A,1)=.975), and reliability of the 30-replication mean exceeded .999 for both models (Table 1). Complementary indices supported the same conclusion (mean pairwise replication correlations $\approx$ .98; split-half reliability $>$ .998). Moreover, agreement between LLMs on per-item mean ratings was also very high (Pearson r = .993, ICC(A,1)=.989), indicating that observed desirability ratings were stable both within and across models.

B Convergent Validity of LLM-based Social Desirability Ratings

Motivation. Our scale construction relies on social desirability estimates s_j obtained from LLM raters (Appendix A) to form desirability-matched forced-choice blocks. A potential threat to validity is that LLM-based desirability judgments might deviate substantially from human norms, in which case desirability matching could be miscalibrated. We therefore conducted an external convergent-validity check by comparing LLM-based ratings to established human social desirability norms for trait words.

Human benchmark. We used the social desirability norms from Britz et al. (2023), which provide mean social desirability ratings (`SOC_MEAN`) for 500 English trait adjectives collected from large human samples. Let $w \in \{1, \ldots, 100\}$ index a random sample of $W = 100$ adjectives drawn without replacement from this database, and let $y_w \in [-3, 3]$ denote the corresponding human mean desirability rating (with -3 = very undesirable and $+3$ = very desirable).

LLM ratings (10 independent repetitions). For each sampled adjective w, we obtained ratings from the same two rater LLMs used in Appendix A (GPT-5 and Gemini 2.5 Pro). Each repetition queried the LLM in a single-turn request (no chat history) with the same normative instruction framing as in Appendix A (rating "how desirable" a trait is for an adult person) and the 9-point integer response scale $\{1, \ldots, 9\}$. Appendix A indicates that the LLM ratings were highly consistent. Therefore, we collected $R = 10$ repetitions per LLM, each time shuffling the 100 words and batching them into four prompts of 25 words, yielding up to $100 \times 10 = 1000$ raw ratings per LLM.

From raw outputs to per-word means. For each LLM l and repetition r, let $x_{wlr} \in \{1, \ldots, 9\}$ denote the parsed desirability rating returned for word w (missing if the response could not be aligned to the prompt after a single refit attempt). To avoid inflating the effective sample size, we first aggregate repetitions at the word level. Specifically, for each word w and LLM l, we compute the per-word mean $\bar{x}_{wl}$ so that each word contributes a single LLM value to the correlation analysis.

Correlation computation. For each LLM l, Table 2 reports (i) the Pearson product–moment correlation between the vectors $\{\bar{x}_{wl}\}_{w=1}^{W}$ and $\{y_w\}_{w=1}^{W}$ and (ii) the Spearman rank correlation between the same vectors. As shown in this table, the LLM-based desirability ratings closely track the human social desirability norms.

C Construction of the Graded Forced-Choice Inventory

Item pool and desirability inputs. We begin from Goldberg's public-domain IPIP Big-Five factor-marker inventory (Goldberg, 1999), which contains 100 statements covering five domains. At this stage, we decided to exclude two voting-related statements from the original item pool and retained $J = 98$ statements. This is because voting behavior is highly susceptible to construct-irrelevant influences—political attitudes/ideology, cross-country institutional differences, and eligibility differences, thereby confounding Big Five measurement. Each statement j is annotated with a Big Five domain label $f(j) \in \{A, C, E, N, O\}$ and a keying sign

Within-LLM consistency across 30 independent replications (100 items)				
LLM	ICC(A,1)	ICC(A,30)	Mean pairwise r	Split-half r
Gemini 2.5 Pro	.980 [.974, .985]	.9993 [.9991, .9995]	.980	.9986 [.9981, .9991]
GPT-5	.975 [.968, .982]	.9992 [.9989, .9994]	.976	.9984 [.9978, .9988]

Between-LLM agreement on per-item mean ratings (100 items)			
Comparison	Pearson r	Spearman ρ	ICC(A,1)
Gemini 2.5 Pro vs. GPT-5	.993 [.989, .995]	.987	.989 [.984, .993]

Table 1: Agreement of observed desirability ratings. ICC(A,1) and ICC(A,30) denote two-way absolute-agreement ICCs for a single replication and the 30-replication mean, respectively (brackets: 95% confidence intervals). Mean pairwise r is the mean off-diagonal correlation among replications across items. Split-half r is based on 2,000 random 15/15 splits of replications (brackets: 2.5/97.5 percentiles).

Rater LLM	Pearson r (95% CI)	Spearman ρ
GPT-5	0.950 [0.927, 0.966]	0.943
Gemini 2.5 Pro	0.950 [0.926, 0.966]	0.947

Table 2: Correlations between per-word mean LLM desirability ratings $\bar{x}_{wl}$ and human mean social desirability ratings y_w from Britz et al. (2023), computed across the sampled $W = 100$ trait adjectives.

$g_j \in \{+1, -1\}$, where $g_j = +1$ denotes a positively keyed item and $g_j = -1$ a negatively keyed item. The social desirability score s_j for each item is the LLM-based estimate described in Appendix A.

Candidate cross-domain pairs. We considered only cross-domain pairs, i.e., pairs of statements that measure different Big Five traits. Let $\mathcal{P} = \{(j, j') : 1 \leq j < j' \leq J,\ f(j) \neq f(j')\}$ be the set of all such unordered pairs, and let $x_{jj'} \in \{0, 1\}$ indicate whether pair $(j, j') \in \mathcal{P}$ is included in the final inventory. For each candidate pair (j, j'), define the absolute desirability gap $\Delta_{jj'} = |s_j - s_{j'}|$ and a mixed-key indicator $g_{jj'} = \mathbb{I}[g_j \neq g_{j'}]$.

Two-stage (lexicographic) mixed-integer optimization. We selected exactly $P = 30$ pairs. To obtain uniformly tight desirability matching, we solved a two-stage mixed-integer program. In the first stage we minimized the maximum within-pair desirability gap by introducing a continuous variable $m \geq 0$:

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

subject to

$$\sum_{(j,j')\in\mathcal{P}} x_{jj'} = P, \tag{8}$$

$$\Delta_{jj'} x_{jj'} \leq m \qquad \forall (j, j') \in \mathcal{P}, \tag{9}$$

$$\sum_{\substack{(j,j')\in\mathcal{P}:\\ j=\ell \text{ or } j'=\ell}} x_{jj'} \leq 1 \qquad \forall \ell \in \{1, \ldots, J\}, \tag{10}$$

$$0.4P \leq \sum_{(j,j')\in\mathcal{P}} g_{jj'} x_{jj'} \leq 0.6P, \tag{11}$$

and the following balance constraints. First, we enforced equal representation of the five domains by requiring each trait to appear exactly $2P/5 = 12$ times:

$$\sum_{(j,j')\in\mathcal{P}} \mathbb{I}[f(j) = t \text{ or } f(j') = t]\, x_{jj'} = 12 \qquad \forall t \in \{A, C, E, N, O\}. \tag{12}$$

Second, we balanced the *domain-pair* composition by requiring each of the ten unordered trait pairs to occur exactly $P/10 = 3$ times:

$$\sum_{(j,j')\in\mathcal{P}} \mathbb{I}[\{f(j), f(j')\} = \{t, t'\}]\, x_{jj'} = 3 \qquad \forall \{t, t'\} \subset \{A, C, E, N, O\},\ t < t'. \tag{13}$$

Finally, to avoid extreme imbalances in positive/negative keying within any domain, let

$$N_{t,+} = \sum_{(j,j')\in\mathcal{P}} \Big(\mathbb{I}[f(j) = t, g_j = +1] + \mathbb{I}[f(j') = t, g_{j'} = +1]\Big) x_{jj'}, \tag{14}$$

$$N_{t,-} = \sum_{(j,j')\in\mathcal{P}} \Big(\mathbb{I}[f(j) = t, g_j = -1] + \mathbb{I}[f(j') = t, g_{j'} = -1]\Big) x_{jj'}, \tag{15}$$

BLK	Trait (key)	Statement	SD	Trait (key)	Statement	SD	$\lvert\Delta SD\rvert$
1	A+	Accept people as they are.	8.75	O+	Enjoy hearing new ideas.	8.75	0.00
2	A+	Am concerned about others.	8.87	C+	Carry out my plans.	8.87	0.00
3	A+	Am easy to satisfy.	6.63	E+	Talk to a lot of different people at parties.	6.65	0.02
4	A+	Believe that others have good intentions.	7.43	O+	Can say things beautifully.	7.62	0.18
5	A-	Contradict others.	3.23	O-	Am not interested in abstract ideas.	3.17	0.07
6	A-	Cut others to pieces.	1.00	N+	Dislike myself.	1.02	0.02
7	A-	Get back at others.	1.03	C-	Shirk my duties.	1.03	0.00
8	A+	Have a good word for everyone.	8.08	N-	Am not easily bothered by things.	8.07	0.02
9	A-	Have a sharp tongue.	2.58	E-	Retreat from others.	2.62	0.03
10	A+	Respect others.	9.00	N-	Remain calm under pressure.	9.00	0.00
11	A-	Suspect hidden motives in others.	2.57	E-	Would describe my experiences as somewhat dull.	2.53	0.03
12	A+	Treat all people equally.	9.00	C+	Complete tasks successfully.	9.00	0.00
13	C+	Am exacting in my work.	7.92	O+	Have a rich vocabulary.	7.80	0.12
14	C+	Do things according to a plan.	7.93	O+	Have a vivid imagination.	7.93	0.00
15	C-	Don't put my mind on the task at hand.	1.65	N+	Often feel blue.	1.63	0.02
16	C-	Don't see things through.	1.52	N+	Am often down in the dumps.	1.50	0.02
17	C-	Find it difficult to get down to work.	1.95	N+	Get stressed out easily.	1.93	0.02
18	C+	Get chores done right away.	8.07	E+	Make friends easily.	8.07	0.00
19	C+	Make plans and stick to them.	8.75	E+	Cheer people up.	8.68	0.07
20	C-	Need a push to get started.	2.78	E-	Find it difficult to approach others.	2.83	0.05
21	C+	Pay attention to details.	8.17	O+	Enjoy thinking about things.	8.10	0.07
22	E+	Am skilled in handling social situations.	8.57	N-	Rarely lose my composure.	8.57	0.00
23	E+	Am the life of the party.	6.47	O+	Enjoy wild flights of fantasy.	6.45	0.02
24	E-	Avoid contacts with others.	2.17	N+	Fear for the worst.	2.17	0.00
25	E-	Don't talk a lot.	4.98	O-	Believe that too much tax money goes to support artists.	4.90	0.08
26	E-	Keep in the background.	4.70	O-	Do not like poetry.	4.63	0.07
27	E-	Keep others at a distance.	2.70	N+	Am filled with doubts about things.	2.72	0.02
28	N+	Worry about things.	2.70	O-	Have difficulty understanding abstract ideas.	2.63	0.07
29	N-	Rarely get irritated.	8.32	O+	Get excited by new ideas.	8.33	0.02
30	N-	Seldom get mad.	8.05	O+	Carry the conversation to a higher level.	8.07	0.02

Table 3: The final 30-block graded forced-choice inventory constructed by two-stage desirability matching. Trait labels: A=Agreeableness, C=Conscientiousness, E=Extraversion, N=Neuroticism, O=Openness to experience. "+"/"–" indicate item keying. SD is the LLM-estimated social desirability score on a 1–9 scale; $|\Delta SD|$ is the absolute within-block difference.

and enforced that each sign accounts for at least 30% of the selected items within each domain:

$$7N_{t,+} \geq 3N_{t,-} \text{ and } 7N_{t,-} \geq 3N_{t,+} \quad \forall t \in \{A, C, E, N, O\}. \tag{16}$$

Let $m^\star$ denote the optimal value of the first-stage problem. In the second stage, we kept the minimax optimum by adding the constraint $m \leq m^\star + \varepsilon$ (with $\varepsilon = 10^{-9}$) and minimized the total squared desirability mismatch:

$$\min_{x,m} \sum_{(j,j')\in\mathcal{P}} (s_j - s_{j'})^2 x_{jj'}$$
$$\text{s.t.} \quad \text{Equations (8)–(16) and } m \leq m^\star + \varepsilon.$$

We implemented this optimization in `R` using `ompr`/`ROI` and solved both stages with the Gurobi solver.

Resulting inventory. The final inventory contains 30 two-statement blocks (60 unique items). The optimized solution attained a maximum within-block desirability gap of 0.18 (mean 0.03, SD 0.04; range [0.00, 0.18]) on the 1–9 scale. Table 3 lists the resulting pairs, together with their LLM-estimated desirability scores.

D Full IRT model specifications

This appendix documents the Bayesian prior specification, identifiability constraints, and Stan implementation details for the IRT models described in Section 3.3. The likelihood components are given in Equations (1)–(2) (Likert GRM) and (3)–(5) (GFC ordinal Thurstonian IRT).

Response units and pooling across LLMs/conditions. For each response format, we fit a single pooled IRT model across all LLMs and both instruction conditions (Section 3.3, "A shared latent metric across LLMs and conditions"). Each row $i \in \{1, \ldots, N\}$ of the response matrix corresponds to one completed questionnaire under a fixed (LLM l, persona r, condition c) combination. Before model fitting, we excluded incomplete response sets (i.e., rows with missing item/pair responses) to keep the likelihood well-defined.

Q-matrix and keying. The design matrix $Q \in \{0, 1\}^{J\times D}$ is one-hot, so that each item loads on exactly one Big Five domain. Reverse-keyed items are handled by a sign vector $g \in \{-1, +1\}^J$, which flips the direction of the item discrimination.

Parameterization. In both the Likert and GFC models we parameterize the signed item discrimination as $a_j = g_j a_j^+$ with $a_j^+ > 0$. In Stan this is implemented by sampling a positive "strength" parameter a_j^+ and multiplying by g_j inside the likelihood. Ordered category thresholds are represented by an `ordered[K-1]` vector, enforcing strict ordering $\kappa_{j1} < \cdots < \kappa_{j,K-1}$ (Likert) and $\kappa_{p1} < \cdots < \kappa_{p,K-1}$ (GFC).

Priors and scale identification. We use weakly informative priors that are identical across the two response formats:

$$\boldsymbol{\theta}_i \sim \mathcal{N}(0, \mathbf{I}_5), \qquad i = 1, \ldots, N, \tag{17}$$

$$a_j^+ \sim \mathcal{N}^+(0, 0.5), \qquad j = 1, \ldots, J, \tag{18}$$

$$\kappa_{j,k} \sim \mathcal{N}(0, 1.5), \qquad j = 1, \ldots, J, \quad k = 1, \ldots, K-1, \tag{19}$$

$$\kappa_{p,k} \sim \mathcal{N}(0, 1.5), \qquad p = 1, \ldots, P, \quad k = 1, \ldots, K-1. \tag{20}$$

Here $\mathcal{N}^+(0, 0.5)$ denotes a half-normal prior (a Normal$(0, 0.5)$ distribution truncated to $a_j^+ > 0$). The multivariate standard normal prior $\boldsymbol{\theta}_i \sim \mathcal{N}(0, \boldsymbol{I}_5)$ fixes the latent location and scale (mean 0, variance 1 for each trait) and assumes prior independence across the five dimensions; we do not estimate a latent correlation matrix among traits in the IRT scoring step.

Link to Stan's `ordered_logistic` parameterization. In Stan we implement both models with `ordered_logistic(`η`, `κ`)`. Stan's ordered-logistic cutpoint convention is $\Pr(Y \le k) = \text{logit}^{-1}(\kappa_k - \eta)$. This is algebraically equivalent to the GRM form in Equation (2) because $\Pr(Y \ge k) = 1 - \Pr(Y \le k-1) = \text{logit}^{-1}(\eta - \kappa_{k-1})$.

GFC scaling convention. Following Equation (4), the GFC linear predictor uses the standardized utility difference $\eta_{ip} = (\mu_{i,R_p} - \mu_{i,L_p})/\sqrt{2}$. In Stan this is implemented by multiplying the right-minus-left difference by a constant $1/\sqrt{2}$ (stored as `inv_sqrt2`). This keeps the a priori scale of the comparison signal comparable to the single-item GRM predictor (a difference of two independent utilities would otherwise have roughly twice the prior variance).

Stan implementation and computation. We fit both models in Stan using `cmdstanr`. To reduce computation, we precompute $\boldsymbol{Q}^\top$ once and form the matrix $\boldsymbol{Z} = \boldsymbol{\Theta}\boldsymbol{Q}^\top$ so that $Z_{ij} = \boldsymbol{q}_j^\top \boldsymbol{\theta}_i$; likelihood contributions are then evaluated item-wise (GRM) or pair-wise (GFC) with vectorization over response units. Both Stan programs contain no `generated quantities`; all derived metrics (SDR indices, correlations, and visualizations) are computed in `R` from posterior summaries.

MCMC settings, diagnostics, and posterior summaries. We ran 4 chains, with 200 warmup iterations and 500 post-warmup iterations per chain, using NUTS with `adapt_delta`=0.95 and `max_treedepth`=12 and a fixed random seed. Sampling was parallelized across chains via `cmdstanr`; to avoid oversubscription we set `OMP_NUM_THREADS=1` (no within-chain OpenMP parallelism). We used Stan defaults for parameter initialization and did not apply thinning. Convergence was monitored using $\widehat{R}$ (threshold 1.01) and effective sample sizes. For downstream analyses we use posterior means of θ as point estimates, denoted $\hat{\theta}$ in the main text.

E Persona generation details

We draw $\boldsymbol{z}_i$ independently from a five-dimensional Gaussian distribution with zero mean and covariance Σ:

$$\boldsymbol{z}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_5(\boldsymbol{\mu}, \Sigma), \qquad \boldsymbol{\mu} = \mathbf{0} \in \mathbb{R}^5. \tag{21}$$

We set Σ to the meta-analytic corrected Big Five intercorrelation matrix reported by van der Linden et al. (2010, Table 2), reordered to match our trait ordering (A, C, E, N, O):

$$\Sigma = \begin{pmatrix} 1.00 & 0.43 & 0.26 & -0.36 & 0.21 \\ 0.43 & 1.00 & 0.29 & -0.43 & 0.20 \\ 0.26 & 0.29 & 1.00 & -0.36 & 0.43 \\ -0.36 & -0.43 & -0.36 & 1.00 & -0.17 \\ 0.21 & 0.20 & 0.43 & -0.17 & 1.00 \end{pmatrix}. \tag{22}$$

F Questionnaire administration prompts

This appendix records the *exact* prompt text used to elicit questionnaire responses, as implemented in our data-collection script. Each API call used a single user message that concatenates a persona prefix (Appendix E), an instruction block (HONEST or FAKE-GOOD), and either Likert statements or GFC statement pairs. Angle-bracket placeholders (e.g., `<STATEMENT>`) indicate run-time substitutions and are *not* literal tokens sent to the models.

F.1 Persona prefix (prepended to every request)

```
YOU ARE THE RESPONDENT.

<TRAIT_SENTENCE_O>
<TRAIT_SENTENCE_C>
<TRAIT_SENTENCE_E>
<TRAIT_SENTENCE_A>
<TRAIT_SENTENCE_N>

Answer all questions AS THIS PERSON would.
```

where each <TRAIT_SENTENCE_*> is a single English sentence describing the persona (e.g., "You are very . . . ."), derived from the sampled ground-truth trait vector (Appendix E).

F.2 Instruction blocks

The instruction block was appended immediately after the persona prefix.

HONEST instruction.

```
You will complete a personality questionnaire.
When completing this questionnaire,
we would like you to be as honest as possible,
that is, to present yourself as you really are.
```

FAKE-GOOD instruction.

```
You will complete a personality questionnaire.
When completing this questionnaire,
we would like you to answer the questions in
order to give a really good impression of
yourself,
that is, to present yourself in the best possible
light.
```

F.3 Likert (7-point) prompt

```
<PERSONA_PREFIX>

<INSTRUCTION_BLOCK>

Indicate how accurately the following statement
describes you
using a 7-point scale:
1: Very Inaccurate
2: Moderately Inaccurate
3: Slightly Inaccurate
4: Neither Accurate nor Inaccurate
5: Slightly Accurate
6: Moderately Accurate
7: Very Accurate
Return ONLY one integer (1-7).
Do not include any other text.
++++
Statement: <STATEMENT>
++++
```

F.4 GFC (7-point bipolar) prompt

```
<PERSONA_PREFIX>

<INSTRUCTION_BLOCK>

For the following pair of statements, indicate
which one describes you
more accurately and by how much
using a 7-point bipolar scale:
1: LEFT statement describes me much more
accurately
2: LEFT statement describes me moderately more
accurately
3: LEFT statement describes me slightly more
accurately
4: About the same
5: RIGHT statement describes me slightly more
accurately
6: RIGHT statement describes me moderately more
accurately
7: RIGHT statement describes me much more
accurately
Return ONLY one integer (1-7).
Do not include any other text.
++++
LEFT: <LEFT_STATEMENT>  ||  RIGHT:
<RIGHT_STATEMENT>
++++
```

Left/right randomization. For each GFC pair, the assignment of the two statements to the LEFT vs. RIGHT slot was randomized; the numeric anchors above therefore always refer to the *displayed* left/right positions.

F.5 Output acceptance and retries

A response was accepted if (i) it contained the expected number of integers for that request and (ii) all integers lay in $\{1, \ldots, 7\}$. Requests failing these checks (including outputs with extra text that prevented extracting the expected count) were re-submitted with the identical prompt up to three additional times; at this point, there were no remaining failures.