Automated Item Neutralization for Non-Cognitive Scales: A Large Language Model Approach to Reducing Social-Desirability Bias

Sirui Wu
University of British Columbia
2125 Main Mall,
Vancouver, BC V6T1Z4, Canada
sirui.wu@ubc.ca

Daijin Yang
Northeastern University
360 Huntington Ave,
Boston, MA 02130, USA
yang.dai@northeastern.edu

Abstract

This study evaluates item neutralization assisted by the large language model (LLM) to reduce social desirability bias in personality assessment. GPT-o3 was used to rewrite the International Personality Item Pool Big Five Measure (IPIP-BFM-50), and 203 participants completed either the original or neutralized form along with the Marlowe–Crowne Social Desirability Scale. The results showed a preserved reliability and a five-factor structure, with gains in conscientiousness and declines in Agreeableness and Openness. The correlations with social desirability decreased for several items, but inconsistently. Configural invariance held, though metric and scalar invariance failed. Findings support AI neutralization as a potential but imperfect bias-reduction method.

1 Introduction

Large language models have primarily been applied to generate cognitive test items and have shown strong performance. With proven powerful contextual understanding and generation abilities in multiple domains (Fitria, 2023; Yang et al., 2025; Ullah et al., 2024), systems such as GPT-3 (Floridi and Chiriatti, 2020) have already produced acceptable multiple choice reading passages (Shin et al., 2025), chemistry and physics items (Chan et al., 2025), and tasks that assess fluid reasoning and visual processing (Ryoo et al., 2022). However, using LLMs for non-cognitive assessments (personality, attitudes, social-emotional skills) is still rare. These constructs are often abstract, value-laden, and context-dependent, which makes automatic item-writing challenging.

Nonetheless, early research is beginning to explore this space. Li et al. (2024a) used GPT-4 to create short, scenario-based questions, named situational judgment items. These items ask people how they would respond in everyday situations, as a way of measuring the Big Five personality traits. In another example, Xue et al. (2025b) relied on GPT-3.5 to expand and translate a university-belongingness questionnaire, maintaining good reliability despite some noisy items. These findings suggest that LLMs can assist non-cognitive scale development, but their robustness and effectiveness remain unverified. Studies have shown that LLM outputs for complex social constructs, such as political or moral values, tend to be overly uniform (Park et al., 2024).

Most prior work has focused on generating new items from scratch, but refining existing validated items through targeted edits is an equally promising yet understudied approach. As emphasized by The Standards for Educational and Psychological Testing (Eignor, 2013), adapting item wording — whether for clarity, cultural context, or bias reduction — can enhance accessibility and fairness while preserving construct validity. McCrae et al. (McCrae et al., 2005) demonstrate that systematic item refinement, like simplifying complex terms in the NEO-PI-3, improves readability and reliability without changing the test's core structure. Studies also show that employing various refinement strategies, such as rephrasing and balancing item tone, can enhance validity while preserving construct discrimination (Bäckström et al., 2014).

To research how LLM could serve as a precise editors, we adopted LLM to identify and decrease the social desirability bias. Social-desirability bias is a tendency for a person to respond in a way that seems socially appealing, regardless of his or her true characteristic (Grimm, 2010; Furr, 2021; Braun et al., 2001). It can contaminate true levels of trait and comparison of individuals, especially on traits such as agreeableness, conscientiousness, and emotional stability (Bäckström and Björklund, 2014), so curbing it is critical. It was chosen for our study not only because it is a common threat to non-cognitive tests, but also because traditional approaches to reducing social desirability, including

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 1–13

forced choice (Cao and Drasgow, 2019), balanced keying (Gignac, 2013; Li et al., 2024c), and manual "neutralization" of wording (Bäckström and Björklund, 2020; Bäckström and Björklund, 2013), can work but are labor intensive and may create unintended dimensions (Zhang et al., 2025).

Recent studies demonstrate that LLMs not only display social desirability response patterns similar to humans, but can also detect when they are being evaluated and shift their answers toward socially valued traits. This ability to recognize and reproduce bias suggests that LLMs could also be leveraged to diagnose and potentially mitigate social desirability effects in human surveys (Lee et al., 2024; Salecha et al., 2024). For instance, Dukanovic and Krpan (2025) conducted a real-world hiring study. They required candidates completed both a standard multiple-choice personality questionnaire and a short conversation with an AI chatbot. The chatbot analyzed their written answers and generated personality scores, and they found chatbot-based scores were less influenced by social desirability than the traditional questionnaire scores. However, the chatbot scores were also less effective at predicting external outcomes such as education level or job role. Nevertheless, few studies have evaluated whether LLMs can rewrite test items to reduce their social desirability without compromising reliability and validity.

To address this gap, we used prompt engineering to guide GPT-o3 in revising the IPIP-BFM-50 (Tao et al., 2009), maintaining the test's structure while reducing social desirability bias. The prompt integrated established debiasing strategies (Kajonius, 2017; Bäckström et al., 2014) and incorporated role-playing (Kong et al., 2023), chain-of-thought prompting (Wei et al., 2022), and transparency mechanisms (Schneider, 2024). We evaluated the AI-neutralized items with participants against the original form, examining reliability, factor structure, and correlations with the Marlowe–Crowne Social Desirability Scale.

The results show that AI-based neutralization attenuated social desirability bias while preserving the Big Five structure within each form. Reliability was maintained in most domains, improved for Conscientiousness, but decreased for Agreeableness and Openness. Confirmatory factor analyses supported configural invariance, though full metric and scalar invariance across versions was not achieved. Correlations with social desirability weakened for many items, though effects were uneven across traits.

The discussion highlights both the promise and limitations of AI-assisted item editing. AI neutralization provides a viable tool for reducing response bias without altering trait constructs, but its uneven performance and lack of cross-form equivalence indicate the need for domain-specific fine-tuning, iterative refinement, and human-in-the-loop validations. Taken together, this work demonstrates the potential of large language models to contribute to fairer psychological assessment through targeted item rewriting.

2 Methods

2.1 Instruments

2.1.1 The International Personality Item Pool Big Five Personality Scale (IPIP-BFM-50)

We employed the IPIP-BFM-50 as the foundational measure of the Big Five personality traits, including 50 items (Goldberg et al., 2006; Zheng et al., 2008). Each personality was measured by 10 items. This version of the IPIP-BFM-50 has been previously validated and shown to retain acceptable psychometric properties across multiple studies (Tao et al., 2009). Across multiple cultural validations, Cronbach's alphas are generally high (.80 – .90) for Extraversion, Conscientiousness, Emotional Stability, and Openness, though Agreeableness is sometimes lower (.65 – .70) (Goldberg et al., 2006; Ypofanti et al., 2015; Zheng et al., 2008). Studies also observed validity evidence based on internal structure and relations to other scales. Factor analyses consistently replicate the expected five-factor structure, with strong invariance across gender and ethnic groups (Constantinescu and Constantinescu, 2016; Buchanan et al., 2005; Ehrhart et al., 2008). Validity is shown through substantial correlations with other Big Five instruments including the NEO Five-Factor Inventory (Gow et al., 2005), the Ten Item Personality Inventory (Ypofanti et al., 2015), and the Eysenck Personality Questionnaire–Revised (Gow et al., 2005), often above .60.

2.1.2 IPIP-BFM-50 with AI-neutralization (IPIP-BFM-50-AI)

To systematically reduce social desirability bias in personality assessments, we developed a tailored prompt for GPT-o3, producing the IPIP-BFM-50-AI. GPT-o3 was selected for its strong instruction

following, long-context reasoning, and coherent, multi-step outputs (Kim et al., 2025; Ballon et al., 2025; OpenAI, 2025). Our design draws on Bäckström et al.'s manual rewriting strategies (Bäckström et al., 2014), emphasizing reduced evaluative language, preserved behavioral meaning, and midpoint-oriented phrasing—methods shown to reduce item popularity while maintaining validity.

Beyond psychometric strategies, the prompt incorporates techniques to boost effectiveness and interpretability. It frames GPT-o3 as an expert psychometrician (Kong et al., 2023), applies chain-of-thought prompting (Wei et al., 2022) to structure reasoning, and enforces transparency through structured outputs with justifications, bias ratings, and fidelity checks (Schneider, 2024). The full prompt and generated items are included in the appendix.

2.1.3 The Marlowe-Crowne Social Desirability Scale (MC-SDS) short form

The SDS is a validated and widely used measure for assessing socially desirable responding. The Marlowe–Crowne Social Desirability Scale (MC-SDS) short forms, particularly the 13-item Reynolds version, exhibit acceptable internal consistency ($\alpha = .76$ (Reynolds, 1982)) and very high correlations (.80–.90 (Ii and Sipps, 1985)) with the full 33-item scale. This evidence supports their reliability and validity.

2.2 Participants and Data Collection

Participants were recruited online through public advertisements and social networks. Eligibility required age 18 or older, and consent to participate. After reading the study information page and providing informed consent, participants were randomly assigned to complete either the original or the AI neutralized version of the IPIP-BFM-50, creating a between-group design with two independent samples. The two forms used identical content domains but different wording where applicable for the AI-neutralized version. To prevent memory and sequence effects, item order was independently randomized within each version, and the version order was counterbalanced across participants. The response format used a 5-point radio-button scale for all items. The Marlowe–Crowne Social Desirability Scale (short form) was administered after one of the two IPIP administrations. Demographic information (age, gender, education, occupation) was collected at the end to minimize priming. All participants were voluntary recruited by an online link, operated and delivered by a free online survey tool Wjx [1]. We collected 203 response, 102 for Original version and 101 for Neutralized version. After excluding all cases with incomplete items, the sample size was finalized to be 200, each version with 100 responses.

2.3 Evaluation Strategy and Hypothesis

2.3.1 Effectiveness of neutralization.

Item and scale level indicators. We will compare item popularity (means, SDs) and scale means between original and AI-neutralized items to check that highly evaluative items show reduced extreme endorsement without loss of variability.

Desirability linkage. Estimate the correlation of each domain with MC-SDS for the original and neutralized versions within persons. Test whether the neutralized version shows a smaller association with MC-SDS.

2.3.2 Validity evidence following the Standards.

Internal structure and reliability. For each version, test unidimensionality within each domain via CFA or IRT dimensionality checks, then test cross-version invariance (configural, metric, scalar) and report reliability (Cronbach alpha).

Relations to other variables. As discriminant evidence, verify that neutralized scales show weaker correlations with social desirability than originals, while preserving expected convergent patterns with established Big Five constructs.

2.3.3 Hypothesis

Results from all analysis above can be used to check the following hypothesis:

1. **H1 - reliability**: Neutralized domains will demonstrate acceptable reliability that is comparable to originals.

2. **H2 - structure**: Each domain will show a single intended factor per version and acceptable cross-version invariance indices.

3. **H3 - relations**: Neutralized domains will maintain expected convergent patterns with Big Five constructs while showing reduced linkage to social desirability.

[1] https://www.wjx.cn/

3 Results

Two balanced groups completed the original and AI-neutralized versions (n = 100 each). As shown in Table 1, most participants were between 26 and 40, and there were also respondents aged from 41 to 50, as well as a small 60+ group. Gender distributions were comparable across versions, with roughly equal numbers of men and women. The groups appear demographically similar, supporting a fair comparison of psychometric results between original and neutralized items.

[Table 1 about here.]

3.1 H1: Reliability

Reliability was largely preserved after neutralization. As shown in Table 2, extraversion and Neuroticism remained high in both versions. Conscientiousness improved in the neutralized form. Agreeableness and Openness decreased, with Agreeableness dropping to the mid $.50s$ – $.60s$. Overall, alpha and omega were acceptable for most domains, indicating that neutralization did not broadly undermine internal consistency, though Agreeableness warrants caution. These findings support H1 with noted caveats.

[Table 2 about here.]

3.2 H2: The Validity Evidence from Internal Structure

Single-group confirmatory factor analyses (CFAs) supported the intended five-factor structure for each version. As shown in Table 3, model fit was acceptable for the original version (CFI $\approx$ .97, TLI $\approx$ .97, RMSEA $\approx$ .06) and marginally weaker for the neutralized version (CFI $\approx$ .97, TLI $\approx$ .96, RMSEA $\approx$ .08). Both versions retain the five-factor structure, but the higher RMSEA in the neutralized form points to a few items needing targeted wording revision.

Multi-group tests showed that configural form held, but metric and scalar constraints produced significant misfit with elevated RMSEA, indicating a lack of full cross-version equivalence. Thus, Hypothesis 2 is partially supported: the structure replicates within versions, but strict invariance across versions was not achieved. Configural invariance was supported, indicating that the neutralized and original versions share the same five-factor pattern and item-to-factor assignments. This shows that neutralization preserved the construct blueprint. However, subsequent metric and scalar constraints did not hold, which implies differences in loadings and intercepts across forms. Scores can be interpreted within each form using the same domain structure, but cross-form comparisons of factor means should be deferred until partial invariance or alignment is applied.

[Table 3 about here.]

3.3 H3: The Validity Evidence from Relations to Other Variables

As shown in Table 4, the results highlight differences in correlations between individual items across the five dimensions and the SDR score. It is expected to observe a decrease in difference for absolute value of correlation (no matter a positive or negative), indicating a decrease of influence by SDR. However, we can observe correlations are increase for some items. We conduct the Steiger's Z test to check whether the change in correlation significant, and 6 items indicate a significant change. Among them, five are decrease and one increase.

The neutralized items demonstrated reduced associations in several cases, supporting the intended effect. However, the presence of increases underscores uneven performance across content. Overall, H3 is partially supported: the tool attenuates social desirability bias for many items, but not consistently across the full instrument.

Table 5 further showed details about what items was assessed to have significant change in correlations with SDR after neutralization. The changes align with specific linguistic mechanisms. For Extraversion, neutralized phrasings replace overt status claims with modest, observable behaviors or internal states. This lowers self-presentational stakes and reduces the incentive to answer in a socially approved way. For Openness, edits remove prestige cues (for example, "rich vocabulary") and normalize difficulty with abstract content. Endorsing these becomes less face-threatening, so links to desirability weaken. The Agreeableness increase arises from hedged, evaluative wording ("others might find rude" and "sometimes"). This introduces norm salience and plausible deniability, inviting impression management more than the blunt behavior label "insult people." In short, SDR decreases when wording is concrete, behavioral, and low in status or virtue signals; SDR increases when wording invokes social judgment, hedges frequency, or allows reframing of intent.

To sum up, AI neutralization works, but not uni-

formly. It maintains reliability in most domains, preserves the factor structure within forms, and reduces desirability in several areas. The costs are local and fixable: a handful of items drive non-invariance and dips in Agreeableness and Openness. Treat scores as within-form for now, apply partial invariance or alignment for cross-form comparisons, and revise the flagged items to restore behavioral precision while keeping neutral tone.

[Table 4 about here.]

[Table 5 about here.]

4 Discussion

The findings indicate that AI-based neutralization can reduce socially desirable responding while preserving the intended construct structure of a Big Five inventory. Single-group CFAs recovered the five-domain pattern in both versions, which suggests that the core representation of the traits remained intact after neutralization. Multi-group analyses supported configural invariance but not metric or scalar invariance, which implies that some item–factor relations and intercepts changed across versions. Reliability remained acceptable for most domains, improved for Conscientiousness, and declined for Agreeableness and Openness. Associations with a social desirability criterion decreased for several items, with notable exceptions in Agreeableness. Together, the results support AI neutralization as a viable wording intervention that targets response bias without altering trait identity.

4.1 Construct representation and measurement comparability

The preserved five-factor structure indicates that neutralization did not shift the meaning of the constructs, which aligns with evidence that the Big Five structure is robust across formats and raters (McCrae and Costa, 1987). The lack of metric and scalar invariance signals that item functioning changed across versions, so cross-form comparisons of means should not be made without partial invariance or alignment solutions (Byrne et al., 1989; Putnick and Bornstein, 2016). Within each form, factors can be interpreted in the usual way. Across forms, unit and intercept differences should be addressed before comparing group or condition means.

4.2 Domain-specific reliability shifts

Conscientiousness reliability increased in the neutralized form, which is consistent with the idea that removing evaluative phrasing can sharpen behavioral focus and raise inter-item coherence. Declines in Agreeableness and Openness suggest that some edits broadened meanings or removed construct-diagnostic cues that previously fostered homogeneity. This pattern is compatible with prior work showing that evaluative wording can inflate internal consistency by cueing a general "goodness" factor, and that neutralizing language can reduce that inflation while leaving substantive variance intact (Bäckström et al., 2014; Bäckström and Björklund, 2013).

4.3 Why correlation with SDR changed

Reductions in correlation with social desirability appear, when wording shifts from status or virtue claims to concrete behaviors or internal states. This likely weakens impression management, which is one facet of socially desirable responding (Paulhus and Reid, 1991). Increases were observed when neutralized items introduced hedges or explicit social judgment cues, which can heighten norm salience and invite self-presentation. These mechanisms align with research on common method bias and evaluative content as drivers of spurious covariance and inflated correlations (Podsakoff et al., 2003; Bäckström and Björklund, 2013).

4.4 Implications for AI-assisted item editing

The results indicate the potential of AI-assisted item editing. Recent research has shown that LLMs themselves exhibit human-like social desirability biases when responding to personality questionnaires, which implies that they are sensitive to the evaluative cues embedded in item wording and may therefore be leveraged to identify and mitigate such bias (Chan et al., 2025). This capacity provides a foundation for the observed reduction in socially desirable responding when items are neutralized with AI support.

However, as the results suggest, a one-time output from a single prompt may not achieve the ideal output. Studies across multiple domains have found that single-shot generation often produces variable quality and is less reliable for tasks requiring precision, nuance, or consistency (Patel et al., 2023; Sahoo et al., 2024). The variability is partly due to the probabilistic nature of LLMs and the dif-

ficulty of capturing subtle linguistic properties in a single attempt. Research on prompting and iterative generation shows that multiple candidates and refinement loops generally outperform one-shot outputs, which supports the interpretation that item editing requires more than a single pass (Cheng et al., 2024; Xue et al., 2025a).

Besides using single prompts, other techniques for enhancing large language model behavior are suggested. For the model itself, domain-specific fine-tuning has been shown to substantially improve performance even when only a small amount of high-quality training data is available (Jeong, 2024; Satterfield et al., 2024). In this context, including pairs of successfully human-edited and neutralized items could increase the model's ability to generate valid revisions. However, such data are difficult to obtain, and constructing this type of dataset is therefore an important future direction.

To add control to the system, multiple agents can be combined to provide feedback and review of generated items. One approach is to use another large language model as a reviewer, which can rate and critique generated items. Generate–feedback loops of this kind have proven effective in other domains, such as reasoning and dialogue, by reinforcing higher quality outputs through self-critique and refinement (Li et al., 2024b; Madaan et al., 2023). Beyond automated feedback, incorporating humans in the loop transforms item generation into an iterative process. In such a cycle, participants test the items, results are analyzed, and the items are further refined based on psychometric evidence. This practice reflects established best practices in test development (Eignor, 2013), where iterative pilot testing and expert review are essential to ensure reliability and validity. Yet, the human-LLM collaboration still remains unexplored in the item editing field.

In summary, the results highlight the potential of AI-assisted item editing but also point to current limitations when relying on single-prompt outputs. Future development will benefit from domain-specific fine-tuning, multi-agent or human-in-the-loop feedback mechanisms, and iterative refinement processes that mirror traditional psychometric standards. Together, these strategies can convert AI neutralization into a reproducible pipeline that reduces bias while maintaining the measurement of intended psychological constructs.

4.5 Limitations

The study used a single language, a single instrument, and a between-groups design in a low-stakes context. Social desirability effects can be stronger under incentives to self-present, which limits generalizability to high-stakes settings. All measures were self-report and collected in one session, which raises the possibility of common method variance despite anonymity instructions. The analyses focused on internal structure, reliability, and associations with a bias criterion, so criterion-related validity with external outcomes remains untested for the neutralized form.

Future work should test neutralized items in high-stakes contexts, use within-person designs to estimate per-respondent reductions in bias, and include informant or behavioral criteria to address common method concerns. Partial invariance searches or alignment should be applied to enable cross-form comparisons, and results should document the number and type of freed parameters. The AI pipeline should be benchmarked across models and prompts, with a reusable library of prompt patterns and failure cases by domain. Replication across languages and populations, test–retest studies, and evaluation of predictive validity will clarify whether bias reduction is achieved without loss of criterion-related information.

5 Conclusion

AI-based neutralization reduced social desirability bias while preserving the Big Five construct structure. Reliability shifts varied across domains, improving for Conscientiousness but declining for Agreeableness and Openness, reflecting the influence of evaluative language on internal consistency. Configural invariance was supported, but metric and scalar invariance were not, indicating that cross-form comparisons require partial invariance or alignment methods. The discussion highlights that AI-assisted item editing is promising but uneven, and future development should emphasize domain-specific fine-tuning, iterative refinement, and human-in-the-loop validation to ensure stable and valid measurement.

References

Martin Bäckström and Fredrik Björklund. 2014. Social desirability in personality inventories. *Journal of Individual Differences*, 35(3):144–157.

Martin Bäckström and Fredrik Björklund. 2020. The properties and utility of less evaluative personality scales: Reduction of social desirability; increase of construct and discriminant validity. *Frontiers in psychology*, 11:560271.

Martin Bäckström, Fredrik Björklund, and Magnus R Larsson. 2014. Criterion validity is maintained when items are evaluatively neutralized: Evidence from a full–scale five–factor model inventory. *European Journal of Personality*, 28(6):620–633.

Marthe Ballon, Andres Algaba, and Vincent Ginis. 2025. The relationship between reasoning and performance in large language models–o3 (mini) thinks harder, not longer. *arXiv preprint arXiv:2502.15631*.

Henry I Braun, Douglas N Jackson, and David E Wiley. 2001. Socially desirable responding: The evolution of a construct. In *The role of constructs in psychological and educational measurement*, pages 61–84. Routledge.

Tom Buchanan, John A Johnson, and Lewis R Goldberg. 2005. Implementing a five-factor personality inventory for use on the internet. *European Journal of Psychological Assessment*, 21(2):115–127.

Barbara M Byrne, Richard J Shavelson, and Bengt Muthén. 1989. Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychological bulletin*, 105(3):456.

Martin Bäckström and Fredrik Björklund. 2013. Social desirability in personality inventories: Symptoms, diagnosis and prescribed cure. *Scandinavian Journal of Psychology*, 54(2):152–159.

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.

Kuang Wen Chan, Farhan Ali, Joonhyeong Park, Kah Shen Brandon Sham, Erdalyn Yeh Thong Tan, Francis Woon Chien Chong, Kun Qian, and Guan Kheng Sze. 2025. Automatic item generation in various STEM subjects using large language model prompting. *Computers and Education: Artificial Intelligence*, 8:100344.

Yu Cheng, Jieshan Chen, Qing Huang, Zhenchang Xing, Xiwei Xu, and Qinghua Lu. 2024. Prompt sapper: A LLM-empowered production tool for building AI chains. *ACM Trans. Softw. Eng. Methodol.*, 33(5).

PM Constantinescu and I Constantinescu. 2016. The adaptation of the big-five IPIP-50 questionnaire in romania revisited. *Bulletin of the Transilvania University of Braşov. Series VII: Social Sciences• Law*, pages 129–138.

Danilo Dukanovic and Dario Krpan. 2025. Comparing chatbots to psychometric tests in hiring: reduced social desirability bias, but lower predictive validity. *Frontiers in Psychology*, 16:1564979.

Karen Holcombe Ehrhart, Scott C Roesch, Mark G Ehrhart, and Britta Kilian. 2008. A test of the factor structure equivalence of the 50-item IPIP five-factor model measure across gender and ethnic groups. *Journal of Personality Assessment*, 90(5):507–516.

Daniel R Eignor. 2013. The standards for educational and psychological testing.

Tira Nur Fitria. 2023. Artificial intelligence (AI) technology in OpenAI ChatGPT application: A review of ChatGPT in writing english essay. In *ELT Forum: Journal of English Language Teaching*, volume 12, pages 44–58.

Luciano Floridi and Massimo Chiriatti. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

R Michael Furr. 2021. *Psychometrics: an introduction*. SAGE publications.

Gilles E Gignac. 2013. Modeling the balanced inventory of desirable responding: Evidence in favor of a revised model of socially desirable responding. *Journal of Personality Assessment*, 95(6):645–656.

Lewis R Goldberg, John A Johnson, Herbert W Eber, Robert Hogan, Michael C Ashton, C Robert Cloninger, and Harrison G Gough. 2006. The international personality item pool and the future of public-domain personality measures. *Journal of Research in personality*, 40(1):84–96.

Alan J Gow, Martha C Whiteman, Alison Pattie, and Ian J Deary. 2005. Goldberg's 'IPIP'big-five factor markers: Internal consistency and concurrent validation in scotland. *Personality and Individual Differences*, 39(2):317–329.

Pamela Grimm. 2010. *Social Desirability Bias*. John Wiley & Sons, Ltd.

Avery Zook Ii and Gary J Sipps. 1985. Cross-validation of a short form of the Marlowe-Crowne Social Desirability Scale. *Journal of Clinical Psychology*, 41(2):236–238.

Cheonsu Jeong. 2024. Fine-tuning and utilization methods of domain-specific LLMs. *arXiv preprint arXiv:2401.02981*.

Petri J Kajonius. 2017. Cross-cultural personality differences between east Asia and northern Europe in IPIP-NEO. *International Journal of Personality Psychology*, 3(1):1–7.

Su Hwan Kim, Severin Schramm, Lena Schmitzer, Kerem Serguen, Sebastian Ziegelmayer, Felix Busch, Alexander Komenda, Marcus Makowski, Lisa C Adams, Keno K Bressem, and 1 others. 2025. Evaluating large language model-generated brain MRI protocols: Performance of GPT-4o, o3-mini, DeepSeek-R1 and Qwen2. 5-72B. *medRxiv*, pages 2025–04.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2023. Better zero-shot reasoning with role-play prompting. *arXiv preprint arXiv:2308.07702*.

Sanguk Lee, Kai-Qi Yang, Tai-Quan Peng, Ruth Heo, and Hui Liu. 2024. Exploring social desirability response bias in large language models: Evidence from GPT-4 simulations. *arXiv preprint arXiv:2410.15442*.

Chang-Jin Li, Jiyuan Zhang, Yun Tang, and Jian Li. 2024a. Automatic item generation for personality situational judgment tests with large language models. *arXiv preprint arXiv:2412.12144*.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024b. LLMs-as-judges: a comprehensive survey on LLMs-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Mengtong Li, Bo Zhang, Lingyue Li, Tianjun Sun, and Anna Brown. 2024c. Mixed-keying or desirability-matching in the construction of forced-choice measures? an empirical investigation and practical recommendations. *Organizational Research Methods*, page 10944281241229784.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Robert R McCrae and Paul T Costa. 1987. Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology*, 52(1):81.

Robert R McCrae, Paul T Costa, Jr, and Thomas A Martin. 2005. The NEO–PI–3: A more readable revised NEO personality inventory. *Journal of personality assessment*, 84(3):261–270.

OpenAI. 2025. OpenAI o3 and o4-mini system card. System card, OpenAI, San Francisco, CA. Version 2 of the Preparedness Framework.

Peter S Park, Philipp Schoenegger, and Chongyang Zhu. 2024. Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods*, 56(6):5754–5770.

Dhavalkumar Patel, Ganesh Raut, Eyal Zimlichman, Satya Narayan Cheetirala, Girish Nadkarni, Benjamin S Glicksberg, Robert Freeman, Prem Timsina, and Eyal Klang. 2023. The limits of prompt engineering in medical problem-solving: a comparative analysis with ChatGPT on calculation based USMLE medical questions. *MedRxiv*, pages 2023–08.

Delroy L Paulhus and Douglas B Reid. 1991. Enhancement and denial in socially desirable responding. *Journal of personality and social psychology*, 60(2):307.

Philip M Podsakoff, Scott B MacKenzie, Jeong-Yeon Lee, and Nathan P Podsakoff. 2003. Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of applied psychology*, 88(5):879.

Diane L Putnick and Marc H Bornstein. 2016. Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental review*, 41:71–90.

William M Reynolds. 1982. Development of reliable and valid short forms of the Marlowe-Crowne social desirability scale. *Journal of clinical psychology*, 38(1):119–125.

Ji Hoon Ryoo, Sunhee Park, Hongwook Suh, Jaehwa Choi, and Jongkyum Kwon. 2022. Development of a new measure of cognitive ability using automatic item generation and its psychometric properties. *Sage Open*, 12(2):21582440221095016.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Aadesh Salecha, Molly E Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H Ungar, and Johannes C Eichstaedt. 2024. Large language models show human-like social desirability biases in survey responses. *arXiv preprint arXiv:2405.06058*.

Nolan Satterfield, Parker Holbrooka, and Thomas Wilcoxa. 2024. Fine-tuning llama with case law data to improve legal domain performance. *OSF Preprints*.

Johannes Schneider. 2024. Explainable generative AI (GenXAI): A survey, conceptualization, and research agenda. *Artificial Intelligence Review*, 57(11):289.

Dongkwang Shin, Jang Ho Lee, and Kyungmin Kim. 2025. An exploratory study on two automated item generators for generating L2 reading test items. *RELC Journal*, page 00336882251326284.

Peng Tao, Dong Guoying, and Stuart Brody. 2009. Preliminary study of a Chinese language short form of the marlowe–crowne social desirability scale. *Psychological reports*, 105(3_suppl):1039–1046.

Ehsan Ullah, Anil Parwani, Mirza Mansoor Baig, and Rajendra Singh. 2024. Challenges and barriers of using large language models (LLM) such as chatgpt for diagnostic medicine with a focus on digital pathology–a recent scoping review. *Diagnostic pathology*, 19(1):43.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Eric Xue, Zeyi Huang, Yuyang Ji, and Haohan Wang. 2025a. Improve: Iterative model pipeline refinement and optimization leveraging llm agents. *arXiv preprint arXiv:2502.18530*.

Mingfeng Xue, Yunting Liu, and HuaXia Xiong. 2025b. Enhancing non-cognitive assessments with GPT: Innovations in item generation and translation for the university belonging questionnaire. In *Proceedings of Large Foundation Models for Educational Assessment*, volume 264 of *Proceedings of Machine Learning Research*, pages 157–172. PMLR.

Daijin Yang, Erica Kleinman, and Casper Harteveld. 2025. GPT for games: An updated scoping review (2020-2024). *IEEE Transactions on Games*, pages 1–16.

Maria Ypofanti, Vasiliki Zisi, Nikolaos Zourbanos, Barbara Mouchtouri, Pothiti Tzanne, Yannis Theodorakis, and Georgios Lyrakos. 2015. Psychometric properties of the international personality item pool big-five personality questionnaire for the greek population. *Health psychology research*, 3(2):2206.

Xijuan Zhang, Muhua Huang, Jessie Sun, and Victoria Savalei. 2025. Improving the measurement of the big five via alternative formats for the bfi-2. *Journal of Personality Assessment*, pages 1–22.

Lijun Zheng, Lewis R Goldberg, Yong Zheng, Yufang Zhao, Yonglong Tang, and Li Liu. 2008. Reliability and concurrent validation of the IPIP big-five factor markers in China: Consistencies in factor structure between internet-obtained heterosexual and homosexual samples. *Personality and individual differences*, 45(7):649–654.

A Appendix

The Prompt for Neutralizing Self-Report Items

You are an expert psychometrician. Your goal is to reword self-report survey items so they measure the intended vocational interest while minimizing social desirability bias.

Social desirability bias is a type of response bias in research where participants tend to answer questions in a way that they believe will be viewed favorably by others, rather than providing completely honest or truthful responses.

Follow these rules:

- Evaluate each item's social desirability bias. Give each item a score within -5 to 5 where 0 represents the lowest social desirability bias, 5 represents positive social desirability bias (people want to choose the item because they think the item is favorable), and -5 represents negative social desirability bias (people do not want to choose the item because they think the item is unfavorable). Keep the item unchanged if its social desirability bias score is in the zone from -1 to 1.
- Think step-by-step — identify value-laden terms, propose alternatives, and self-check that the new wording still reflects the original behaviour, and that the new wording reduces the social desirability bias — but **do not reveal your reasoning**.
- Remove or soften status-, value-, or social desirability-laden words.
- Construct an item that you would find less desirable yourself.
- If the adjective is evaluatively positive, use a less evaluative one, or rephrase in a way that makes the adjective less evaluative.
- Do not change an item from positive to negative (direction).
- Think of whether the item is reversed or not.
- Preserve each item's core behavioural meaning.
- Pay attention to the dimension of each statement. Do **NOT** change the dimension of each statement.
- Explain your change in natural language for each statement, and give your change a score to indicate its new social desirability bias.

Output format:
Please output the results in a 5-column table titled **Neutralized Items**, with the following headers:

Original | SD Score | Neutralized | Reason | SD Score

Each statement is tagged with a dimension based on the Big Five personality traits. Use the following codes:

- A: Agreeableness
- C: Conscientiousness
- N: Neuroticism

- O: Openness to Experience
- E: Extraversion

The sign "+" or "–" indicates whether the item is positively or negatively phrased within that dimension.

List of Tables

Table 1: Socio-demographics Variable

Age	Original count	Neutralized count
18–25	9	13
26–30	17	24
31–35	59	52
36–40	8	6
41–50	6	3
Over 60	1	2
Gender		
Male	49	46
Female	51	54

Table 2: Reliability for All Subscales

	Original		Neutralized	
Subscale	**Alpha**	**Omega**	**Alpha**	**Omega**
Extraversion	0.90	0.91	0.87	0.89
Agreeableness	0.67	0.71	0.59	0.63
Conscientiousness	0.73	0.77	0.79	0.81
Neuroticism	0.91	0.91	0.94	0.94
Openness	0.78	0.78	0.66	0.71

Table 3: Confirmatory Factor Analysis Model Fit on the Big Five Personality Model

Single-group CFA fit

Group	χ^2 (scaled)	**df**	**p (scaled)**	**CFI**	**TLI**	**RMSEA [90% CI]**	**SRMR**
Original	1284.957	1165	<0.001	0.972	0.971	0.060 [0.033, 0.079]	0.093
Neutralized	1336.244	1165	<0.001	0.965	0.963	0.078 [0.055, 0.097]	0.101

Multi-group invariance

Model	**Df**	**AIC**	**BIC**	χ^2	Δdf	$\Delta\chi^2$	**p**
Configural	2330	25254	26309	4576.3	—	—	—
Metric	2375	25361	26268	4773.8	45	197.51	<0.001
Scalar	2420	25624	26383	5127.0	45	353.16	<0.001

Table 4: Difference in the Correlations with the SDR

	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
delta	-0.14	0.06	-0.11	0.12	0.00
p	0.26	0.66	0.43	0.34	0.98
delta	-0.09	-0.03	-0.01	0.11	-0.29
p	0.47	0.84	0.96	0.41	0.03*
delta	0.16	0.28	0.07	0.04	-0.29
p	0.19	0.03*	0.58	0.77	0.04*
delta	-0.45	0.10	0.03	0.16	-0.03
p	<0.001	0.49	0.83	0.20	0.81
delta	-0.12	-0.12	0.10	-0.13	0.21
p	0.37	0.36	0.41	0.29	0.14
delta	-0.13	0.00	-0.05	0.14	-0.02
p	0.29	0.98	0.71	0.24	0.86
delta	-0.14	-0.11	-0.05	0.06	0.06
p	0.28	0.37	0.72	0.60	0.62
delta	-0.24	-0.09	0.09	0.18	-0.07
p	0.05*	0.53	0.45	0.16	0.64
delta	-0.07	0.08	0.18	0.05	-0.26
p	0.58	0.58	0.20	0.68	0.06
delta	0.10	0.01	0.00	0.04	0.14
p	0.44	0.94	0.98	0.73	0.28

Table 5: Original and neutralized items with SDR correlation changes

Version	**Items**	**Personality**	**Direction**	**Correlation with SDR after neutralization**
Original	Don't mind being the center of attention.	Extraversion	Positive	
Neutralized	Feel fine when attention is on me.	Extraversion	Positive	Decrease
Original	Am the life of the party.	Extraversion	Positive	
Neutralized	Often take an active role in group conversations.	Extraversion	Positive	Decrease
Original	Insult people.	Agreeableness	Negative	
Neutralized	Sometimes say things that others might find rude.	Agreeableness	Negative	Increase
Original	Have a rich vocabulary.	Openness	Positive	
Neutralized	Know and use a variety of words.	Openness	Positive	Decrease
Original	Have difficulty understanding abstract ideas.	Openness	Negative	
Neutralized	Find abstract ideas challenging.	Openness	Negative	Decrease