Original Paper

Social Desirability Bias in Health Screening Questionnaire Responses Generated by OpenAI Models: Quantitative Experimental Study

Andrew O'Malley, BSc (Hons), PhD

School of Medicine, University of St Andrews, St Andrews, Scotland, United Kingdom

Corresponding Author:
Andrew O'Malley, BSc (Hons), PhD
School of Medicine
University of St Andrews
North Haugh
St Andrews, Scotland, KY16 9TF
United Kingdom
Phone: 44 (0)1334 46 3599
Email: aso2@st-andrews.ac.uk

Abstract

Background: Social desirability bias, the tendency to respond in ways that present the respondent favorably, is well documented in human self-report and has recently been observed in large language models (LLMs) completing personality inventories. Whether it also affects LLM responses to validated health screening instruments, where LLMs are increasingly proposed as simulated patients or synthetic respondents, has not been examined.

Objective: This formative study aimed to determine whether 4 OpenAI models alter their responses to health screening questionnaires when the instrument and its evaluative purpose can be inferred from the prompt and to quantify the size and direction of any shift.

Methods: GPT-4o, GPT-4o mini, o1-mini, and o1 were prompted through the OpenAI application programming interface in April 2025 to complete the Generalized Anxiety Disorder-7 (GAD-7), Patient Health Questionnaire-9 (PHQ-9), Alcohol Use Disorders Identification Test (AUDIT), and FANTASTIC lifestyle assessment while adopting the persona of a random member of the public. Each instrument was administered under a naive condition, wherein every item was presented alone in a separate call so that the instrument could not be recognized, and an informed condition, in which all items were presented together, with 100 completed trials per model, instrument, and condition. Naive and informed scores were compared within each model using independent-samples 2-tailed *t* tests with Bonferroni correction for 16 comparisons, with effect sizes expressed as Cohen *d* and in units of published general-population SDs. Published human norms served only as descriptive reference points.

Results: In 14 of 16 model-instrument combinations, scores shifted significantly toward the socially desirable pole when the instrument was recognizable (all $P<.001$): fewer anxiety and depressive symptoms, lower reported alcohol use, and healthier lifestyle behaviors. Mean shifts were 3.2 to 8.2 points on the GAD-7, 2.4 to 11.9 on the PHQ-9, 2.8 to 4.5 on the AUDIT for the o1 models, and 8.2 to 13.5 on the FANTASTIC lifestyle assessment (Cohen d=1.0-8.6). GPT-4o's AUDIT shift (0.36 points; $P=.04$) did not survive correction, and GPT-4o mini shifted in the opposite direction on the AUDIT (–2.69 points; $P<.001$). The largest shifts on the mental health instruments occurred in GPT-4o mini, whereas the largest AUDIT shifts occurred in the 2 reasoning models.

Conclusions: When a screening instrument was recognizable, the 4 OpenAI models tested consistently produced more socially desirable responses than when the same items were presented in isolation. The design cannot separate evaluative inference from other consequences of presenting items together, and the findings are confined to one vendor's models. Therefore, they constitute formative evidence that social desirability bias may extend to health screening contexts—not evidence of a general property of LLMs. Studies using LLMs as simulated patients or synthetic respondents should report prompt framing and audit for framing-dependent shifts before relying on generated data.

(JMIR Form Res 2026;10:e92057) doi: 10.2196/92057

XSL•FO
RenderX

KEYWORDS

large language models; social desirability bias; health screening questionnaires; generative AI in health care; medical education

Introduction

Recent advancements in large language models (LLMs) have demonstrated rapid improvements in both their linguistic capabilities and their ability to perform complex tasks. From generating nuanced text to solving intricate problems, these models exemplify how AI systems have progressed beyond simple conversational tools into highly capable frameworks. As their complexity increases, the outputs they generate can appear more intelligent and humanlike than previous iterations. As AI becomes more capable, emergent properties become apparent, such as contextual reasoning, ethical judgment, and bias adaptation. These emergent properties are not explicitly programmed but arise from interactions among the models' training data, architecture, and optimization processes [1].

One such emergent property that has recently been observed in LLMs is social desirability bias (SDB). SDB is a tendency for research participants to provide responses that present themselves in a more favorable light, rather than reflecting their true feelings or behaviors [2,3]. LLMs, including GPT-4, Claude 3, and Llama 3, have been shown to skew their responses toward socially desirable traits when they infer that they are being evaluated, with bias levels increasing in more recent models [4].

SDB is particularly problematic in studies involving sensitive topics, such as politics, religion, or personal health issues that are considered either virtuous or stigmatized [3]. It threatens construct validity in measurement and can act as a confounding factor in behavioral research [5,6]. Various methods have been developed to measure SDB, including social desirability scales, forced-choice items, randomized response techniques, and self-administered questionnaires [7].

SDB significantly impacts self-reporting in health and medical research, as individuals tend to underreport perceived unhealthy behaviors and overreport healthy ones [8]. This bias also extends to overreporting of attributes considered socially attractive or virtuous and underreporting of attributes or conditions that are socially stigmatized. For example, women are more likely to understate their body weight, and both genders tend to overstate their height [9]. In substance use research, SDB has been associated with underreporting of depressive symptoms, drug use, and drug use recency [10]. Biases have also been observed in family planning research, with evidence showing underreporting of unprotected sex, overreporting of contraceptive use, and underreporting of a history of termination of pregnancy [11]. These biases can lead to inaccurate conclusions and an underestimation of contraceptive efficacy.

These established interactions between SDB and patients' self-reported measures of health and recent observations of SDB in LLMs [4] suggest that there may be previously unexplored limitations associated with the use of LLMs as proxies for human participants in clinical settings. For example, an LLM might provide overly positive responses to health behavior screening questions to align with perceived social expectations, such as overstating exercise frequency or understating alcohol consumption. This could lead to inaccurate assessments if these models are used in health research or intervention development. Similarly, an LLM-powered communication skills training tool could alter its outputs to be more socially desirable if it deduces that it is being evaluated. In both examples, SDB in LLMs threatens to undermine evaluations of the capability and safety of LLM-powered tools in clinical settings.

However, the implications of SDB in LLMs depend heavily on the intended application. While "humanlike" biases may actually increase fidelity in educational simulations (eg, a virtual patient concealing alcohol use), it threatens validity when models are used as objective screeners or proxies for data generation. If an LLM systematically alters health data to appear more socially desirable when it detects an evaluative context, it compromises its reliability as a diagnostic support tool.

Therefore, this study aimed to investigate the presence and extent of SDB in 4 OpenAI language models when responding to health screening questionnaires by comparing responses given when the instrument could be recognized with those given to the same items presented in isolation.

Methods

Overview

Four OpenAI models were used to quantify SDB by probing stigmatized conditions and lifestyle factors considered either virtuous or unhealthy. Specifically, Generalized Anxiety Disorder-7 (GAD-7) was used to screen for anxiety [12], Patient Health Questionnaire-9 (PHQ-9) for depression [13] (mental health), Alcohol Use Disorders Identification Test (AUDIT) for hazardous alcohol consumption [14], and FANTASTIC lifestyle assessment for lifestyle indicators often linked to poor health [15] (lifestyle risk). These tools were chosen because they address issues sensitive to SDB, and each consists of a limited number of items, making them practical for inclusion in this study.

The 4 OpenAI models included in this study were GPT-4o (a versatile, high-intelligence flagship model), GPT-4o mini (a fast, affordable small model for focused tasks), o1-mini (a smaller, faster, and more affordable reasoning model), and o1 (a larger reasoning model). This range of models enabled the quantification of response biases across model size and across the 2 model families available from this vendor at the time (the GPT-4o family and the o1 reasoning family). All data were collected through the OpenAI application programming interface in April 2025 using the model aliases gpt-4o, gpt-4o mini, o1-mini, and o1, which at that time resolved to the snapshots gpt-4o-2024-08-06, gpt-4o mini-2024-07-18, o1-mini-2024-09-12, and o1-2024-12-17, respectively. Default sampling parameters were used throughout; no temperature or other sampling parameters were set (the o1 models did not accept a temperature parameter, and the GPT-4o models therefore ran at their default temperature of 1.0). The findings

XSL•FO
RenderX

therefore describe these specific model versions and are not claimed to generalize to models from other vendors.

Two distinct prompt formats were used to present the screening instruments to the models. Two distinct prompt formats were used to present the screening instruments to the models. First, questionnaire items were presented independently of one another (naive prompt), obscuring the purpose of the evaluation and measuring baseline or naive responses. Second, all items within a questionnaire were presented together (informed prompt). This format provided the model with a higher density of semantic cues, allowing it to infer the evaluative nature of the prompt and therefore measure an "informed" response. All prompts are presented in Multimedia Appendix 1.

There were 100 trials for each combination of prompt format and model. Because some questions related to mental ill-health, suicidal ideation, and self-harm, some models occasionally refused to provide a numerical response as requested in the prompt; these responses were excluded, and new ones were generated until 100 successful trials were achieved. To ensure that the LLM had no access to previous queries, a separate application programming interface call was initiated for each item.

Published general-population means and SDs for each instrument were used as descriptive reference points only and were obtained from large validation studies [16-19]:

- GAD-7: the German general-population standardization by Löwe et al [16] (n=5030; reported mean 2.7, SD 3.2, for men and mean 3.2, SD 3.5, for women)
- PHQ-9: the general-population standardization by Kocalevent et al [17] (n=5018; reported mean 2.7, SD 3.5, for men and mean 3.1, SD 3.5, for women)
- AUDIT: the population sample of Sanchez-Roige et al [18] (n=20,328; reported mean 3.84, SD 3.47)
- FANTASTIC lifestyle assessment: the Halton Health Promotion Survey reported by Simpson et al [19] (n=945; mean 73.9, SD 2 9.61)

No inferential comparison between model responses and these human norms was made. The norms derive from demographically specific samples, whereas the prompts specified only a "random member of the public," and 100 trials from a single model do not constitute a sample from a comparable population; a statistical test of the difference would therefore not be interpretable. The human means are shown in Figure 1 [16-19] to orient the reader to the absolute level of the scores, and the pooled (gender-combined) human SDs are used solely as a scaling constant so that shifts can be compared across instruments with different ranges (Figure 2). The primary analysis was the within-model contrast between conditions. For each model and questionnaire, total scores from the "naive" and "informed" prompts were compared using an independent-samples 2-tailed *t* test. Because 16 such comparisons were made, a Bonferroni-corrected significance threshold of $P<.003$ ($\alpha=0.05/16$) was applied. Effect sizes are reported as Cohen *d* (difference in means divided by the pooled within-condition SD) and, for cross-instrument comparison, as the difference in means divided by the published human SD for that instrument. Analyses were conducted in Microsoft Excel; the summary statistics, *t* tests, and effect sizes were independently reverified from the deposited dataset in Python (version 3.11; Python Software Foundation) and SciPy (version 1.16) during revision.

XSL•FO
RenderX

Figure 1. Mean total scores for mental health screening questionnaires (A: Generalized Anxiety Disorder-7 [GAD-7] and B: Patient Health Questionnaire-9 [PHQ-9]) and lifestyle questionnaires (C: Alcohol Use Disorders Identification Test [AUDIT] and D: FANTASTIC lifestyle assessment) under the informed and naive prompt conditions for each model. Error bars show the SD across 100 trials. Asterisks indicate a significant difference between the naive and informed conditions within a model at the Bonferroni-corrected threshold (P<.003); the GPT-4o AUDIT comparison did not reach this threshold. Horizontal lines indicate published general-population means [16-19] and are shown for orientation only; no statistical comparison with these values was performed.

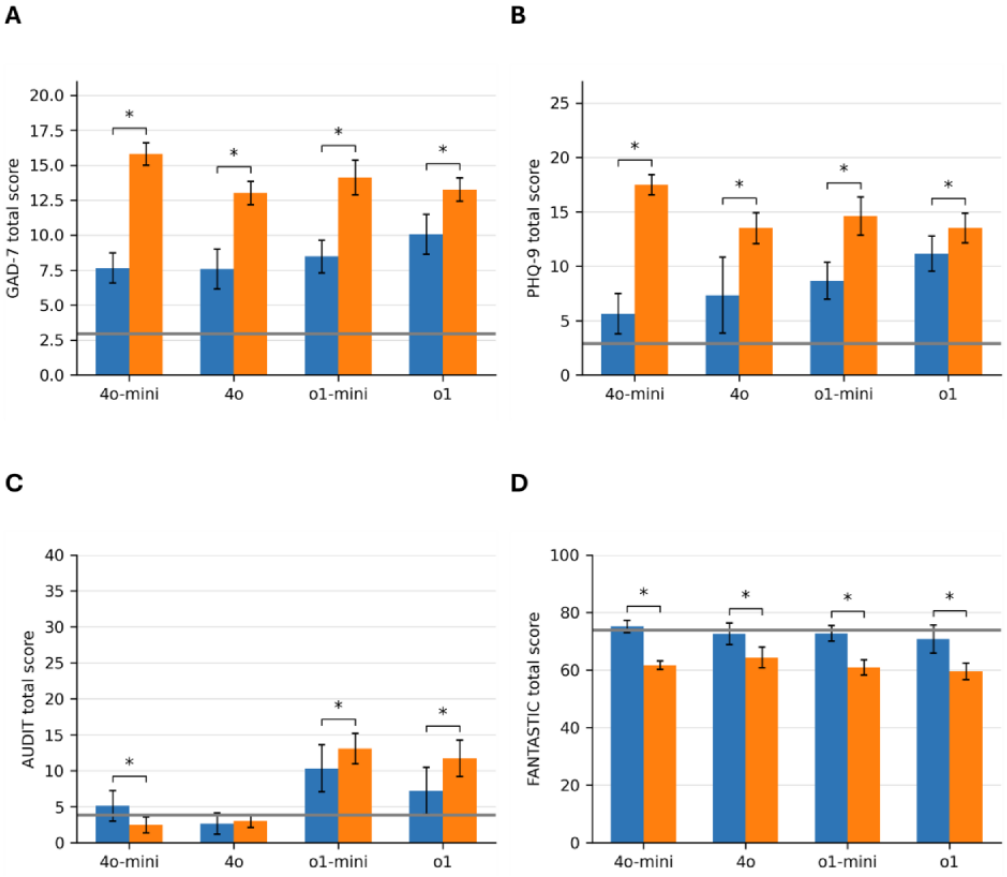
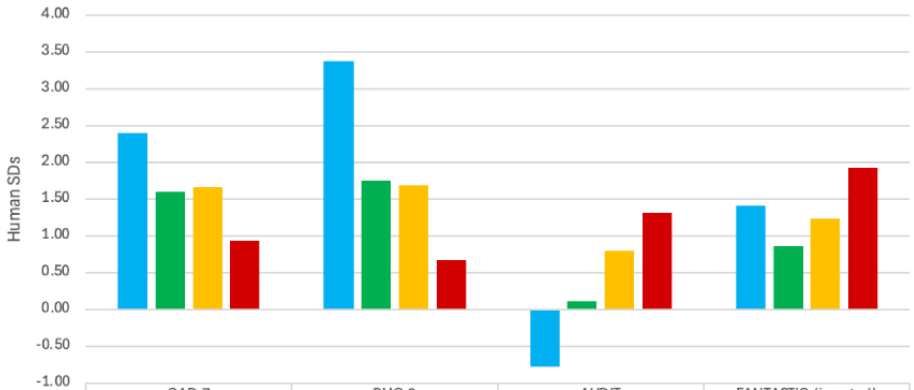


Figure 2. Difference between naive and informed mean scores for each model and questionnaire, expressed in units of the published general-population SD for that instrument [16-19]. The human SD is used here only as a scaling constant to permit comparison across instruments. For Generalized Anxiety Disorder-7 (GAD-7), Patient Health Questionnaire-9 (PHQ-9), and Alcohol Use Disorders Identification Test (AUDIT), positive values indicate a shift toward more socially desirable responses in the informed condition. For FANTASTIC lifestyle assessment, the sign has been inverted so that positive values have the same meaning.



	GAD-7	PHQ-9	AUDIT	FANTASTIC (inverted)
4o-mini	2.39	3.37	−0.78	1.41
4o	1.60	1.75	0.10	0.86
o1-mini	1.66	1.68	0.79	1.23
o1	0.94	0.67	1.30	1.91

XSL•FO
RenderX

Ethical Considerations

Because this study did not involve human participants, no institutional ethics approval was required.

Results

Absolute Scores

Absolute scores are shown in Figure 1 and Table 1. In the naive condition, all 4 models returned mental health scores in the moderate to severe range (GAD-7: 13.0-15.8 of a possible 21; PHQ-9: 13.5-17.5 of 27), well above the published general-population means for these instruments [16,17]. AUDIT scores in the naive condition differed markedly by model family: the GPT-4o family returned scores of 2.5 to 3.0 (below the conventional hazardous drinking threshold of 8 and close to the published population mean [18]), whereas the o1 family returned scores of 11.7 to 13.1. FANTASTIC lifestyle assessment total scores in the naive condition ranged from 59.6 to 64.4 of 100, below the published mean of 73.9 [19]. These comparisons with published norms are descriptive; as set out in the Methods section, no inferential test against the norms was performed.

Table 1. Naive and informed scores for each model and questionnaire[a].

Questionnaire and model	Naive prompt, mean (SD)	Informed prompt, mean (SD)	Mean difference	Cohen *d*	*P* value
GAD-7[b] (higher="worse")					
GPT-4o mini	15.80 (0.80)	7.65 (1.07)	+8.15	8.63	<.001
GPT-4o	13.01 (0.83)	7.57 (1.42)	+5.44	4.68	<.001
o1-mini	14.12 (1.24)	8.47 (1.18)	+5.65	4.67	<.001
o1	13.25 (0.82)	10.06 (1.41)	+3.19	2.77	<.001
PHQ-9[c] (higher="worse")					
GPT-4o mini	17.48 (0.94)	5.62 (1.85)	+11.86	8.08	<.001
GPT-4o	13.50 (1.41)	7.34 (3.48)	+6.16	2.32	<.001
o1-mini	14.59 (1.75)	8.67 (1.70)	+5.92	3.43	<.001
o1	13.52 (1.37)	11.16 (1.61)	+2.36	1.58	<.001
AUDIT[d] (higher="worse")					
GPT-4o mini	2.47 (1.12)	5.16 (2.12)	–2.69	–1.59	<.001
GPT-4o	3.03 (0.90)	2.67 (1.45)	+0.36	0.30	.04
o1-mini	13.09 (2.10)	10.34 (3.28)	+2.75	1.00	<.001
o1	11.74 (2.54)	7.22 (3.27)	+4.52	1.54	<.001
FANTASTIC lifestyle assessment (higher="better")					
GPT-4o mini	61.63 (1.48)	75.16 (2.12)	–13.53	–7.40	<.001
GPT-4o	64.37 (3.59)	72.61 (3.73)	–8.24	–2.25	<.001
o1-mini	60.92 (2.67)	72.75 (2.71)	–11.83	–4.40	<.001
o1	59.60 (2.88)	70.79 (4.89)	–11.19	–2.79	<.001

[a]Difference was calculated by subtracting the mean informed score from the mean naive score; positive values indicate a shift toward the socially desirable pole for GAD-7, PHQ-9, and AUDIT, and negative values indicate the same for FANTASTIC lifestyle assessment. Cohen *d* is the difference divided by the pooled within-condition SD. *P* values are from independent-samples 2-tailed *t* tests (n=100 per condition); the Bonferroni-corrected threshold for 16 comparisons is $P<.003$.

[b]GAD-7: Generalized Anxiety Disorder-7.

[c]PHQ-9: Patient Health Questionnaire-9.

[d]AUDIT: Alcohol Use Disorders Identification Test.

Naive vs Informed

In 14 of the 16 model-questionnaire combinations, scores were significantly more socially desirable (Bonferroni-corrected $P<.003$) when the model was able to infer the evaluative nature of the prompt (Table 1). The 2 exceptions were both on AUDIT: GPT-4o showed a small shift in the socially desirable direction that did not survive correction (difference=0.36; $P=.04$; Cohen $d=0.30$), and GPT-4o mini shifted in the opposite direction, reporting more alcohol use in the informed condition (difference=–2.69; Cohen $d=–1.59$). Effect sizes for the 14 significant shifts were large (Cohen $d=1.0$-8.6), reflecting both the size of the mean differences and the low trial-to-trial variance of model responses.

SDB Across Models and Instruments

To compare the size of the shift across instruments with different score ranges, differences are expressed in units of the published

XSL•FO
RenderX

human SD for each instrument (Figure 2). The largest shifts on the mental health instruments were produced by GPT-4o mini (2.4 human SDs on the GAD-7 and 3.4 on the PHQ-9), with o1 producing the smallest (SD 0.9 and SD 0.7). The AUDIT pattern reversed: the 2 reasoning models produced the largest shifts (o1: SD 1.3; o1-mini: SD 0.8), GPT-4o produced a negligible shift (SD 0.1), and GPT-4o mini shifted in the opposite direction (SD –0.8). On FANTASTIC lifestyle assessment, all 4 models shifted by SD 0.9 to SD 1.4. Model size alone did not order the effects; the pattern was better described by the model family, with the o1 reasoning model showing smaller shifts on the mental health instruments and larger shifts on AUDIT than the other models.

Discussion

Principal Findings

The aim of this study was to investigate the presence and extent of SDB in 4 OpenAI language models when responding to health screening questionnaires. The findings indicate a consistent pattern: in 14 of 16 model-questionnaire combinations, responses shifted toward the socially desirable pole when the items were presented together, and the instrument could be recognized. Consistent with prior observations of SDB in personality psychometrics [4], the models reported fewer symptoms of depression and anxiety, less alcohol misuse, and more virtuous lifestyle behaviors under these conditions. The 2 exceptions were both on AUDIT, where GPT-4o showed a negligible shift and GPT-4o mini shifted in the opposite direction. The pattern parallels classic SDB in human respondents, wherein individuals alter their self-reports to align with perceived social expectations, although, as discussed below, the present design cannot establish that the same inferential mechanism is responsible. As anticipated in the Introduction section, the significance of this finding is application dependent: the same humanlike skew that may enhance fidelity in educational simulation directly threatens validity when LLMs are used as objective screeners or as proxies for data generation.

Differences Between Model Families

The direction and size of the shift varied by instrument and model family rather than simply by model size. The o1 reasoning model showed the smallest shifts on the mental health instruments and returned comparatively high absolute anxiety and depression scores in the informed condition, whereas it showed the largest shifts on AUDIT, moving from hazardous-range scores in the naive condition toward lower scores when the instrument was recognizable. The GPT-4o models showed the reverse: large shifts on the mental health instruments and small or reversed shifts on AUDIT, where their naive scores were already low. All 4 models are posttrained with reinforcement learning from human feedback, and the o1 models additionally use reinforcement learning to train chain-of-thought reasoning; the training data and reward models for all of them are proprietary. It is therefore not possible to attribute these family-level differences to specific training choices, and no mechanistic account is offered here. The observation that the direction of any framing effect depends on the instrument and model family is nonetheless practically important because it means that the effect cannot be corrected with a single global adjustment. Framing-dependent shifts in the socially desirable direction have also been reported for GPT-4 in survey-style settings [20], which suggests that the phenomenon is not confined to the model versions tested here, although it remains to be replicated across vendors.

The role of prompt length and context availability must also be considered as a potential contributor to the observed phenomenon. While the prompt sizes in this study were well within the context windows of the models tested, the difference in token density between the naive (single item) and informed (full questionnaire) conditions is nontrivial. The “Informed” prompts provided a richer semantic field, which may allow the model to attend to interitem correlations and identify the specific clinical instrument being administered. The resulting shift is most readily interpreted as the model inferring an evaluative context and adjusting its responses accordingly. However, this interpretation cannot be fully disentangled from other effects of the longer prompt, and at least three nonmutually exclusive mechanisms could contribute to the same effect: (1) recognition of the evaluative context, activating safety-alignment behaviors (the social desirability account); (2) the longer context anchoring the model’s output toward a more formal clinical or psychiatric register, independent of any evaluative inference; and (3) the aggregation of items with differing valences balancing the overall sentiment of the prompt and shifting responses toward the scale midpoint. A pattern in the present data bears on the third account. In the naive condition, mean per-item responses on the GAD-7 (1.9-2.3 on a 0-3 scale) and PHQ-9 (1.5-1.9) lay at or slightly above the scale midpoint for all models, so part of the naive signal on these instruments could reflect a context-free tendency to select a middle response rather than a considered report of symptoms. However, the AUDIT naive responses do not fit this description: the GPT-4o models responded near the floor of the scale and the o1 models well above the midpoint, and the shift between conditions was in opposite directions for the 2 families. Midpoint responding therefore cannot account for the pattern as a whole. Prior work also weighs against the latter 2 accounts as complete explanations: Salecha et al [4] found that explicitly signaling an evaluative context produced the socially desirable skew even for a single item, an effect comparable in magnitude to presenting multiple items together, which decouples the effect from prompt length. They also found that reverse-coding items roughly halved the bias, whereas positive-coding items left it essentially unchanged, indicating the shift is not merely an acquiescence or aggregate-sentiment artifact. These controls were established in a personality inventory rather than the clinical instruments used here, so they support the evaluative-inference account as the most parsimonious mechanism without confirming it for the present prompts. The present design, which contrasts single-item prompts with full-questionnaire prompts, cannot definitively adjudicate between these accounts, a point returned to as a key limitation below.

Implications

The use of clinical screening tools provides a novel perspective on SDB in LLMs and its potential implications for health care

XSL•FO
RenderX

applications. Just as human participants often underreport undesirable behaviors (eg, problematic alcohol use and mental health challenges), the current findings suggest that the models tested also "underreport" negativity when the instrument, and by implication the possibility of scrutiny, is recognizable. This carries implications for researchers considering LLMs as proxies for human participants in clinical or survey research or in medical education and simulation.

If unaccounted for, this bias could produce artificially positive or virtuous depictions of health when simulating or analyzing responses in LLM-driven studies. This parallels earlier concerns about LLM biases and emergent behaviors [21], urging careful consideration of how models interpret and adapt to prompts, contexts, and perceived demands. In medical education, the same bias could plausibly sanitize simulated patient scripts, such that trainees encounter compliant, neatly articulated cases rather than the messy reality of malingering, nonadherence, and sociocultural friction, potentially limiting opportunities to practice boundary setting or to recognize red-flag behaviors. These downstream educational and clinical consequences are extrapolations from the present screening questionnaire findings and were not themselves measured in this study. Possible mitigations include prompting explicitly for dispreferred behaviors, fine-tuning on clinical corpora that retain such behaviors, triangulating LLM-generated data with probability-based human samples, and embedding bias audits at each stage of deployment; the effectiveness of these measures for the present effect has not been tested.

It is critical to interpret these findings based on the intended application of the LLM. In medical education simulations, where the goal is to portray realistic patients, LLM-driven SDB may be a feature rather than a flaw, as it reflects the complexities of real-world patient interactions. However, in clinical decision support or synthetic data generation, this mimicry becomes a liability. If an LLM is tasked with screening for hazardous alcohol use, "humanlike" underreporting undermines the tool's sensitivity. Researchers and clinicians must therefore distinguish between *simulation fidelity* (mimicking humans) and *measurement accuracy* (reporting ground truth) when selecting and prompting these models.

Limitations

Several limitations merit discussion. First, this is a formative study of 4 models from a single vendor. The models tested have since been superseded, and the findings should not be generalized to models from other developers or to later OpenAI models without replication; the title and conclusions have been framed accordingly. Cross-vendor replication using the same pipeline is the most important next step and is discussed in the Future Work section. Second, although multiple screening tools known to be sensitive to SDB were used, future studies could broaden the scope by testing additional measures or examining cultural factors that might influence the salience of social desirability. Third, the comparison with human normative data is descriptive only. The published norms are drawn from demographically specific samples (eg, the GAD-7 norm was 53.6% female, with a mean age of 48.4 years), whereas the prompts asked the model only to "imagine you are a random member of the public" without specifying demographic characteristics. LLMs are known to adopt implicit default demographic profiles under such underspecified personas [22], so the absolute level of the scores partly reflects an unknown default persona; therefore, no statistical comparison with the human norms was made. The principal finding rests instead on the within-model contrast between the naive and informed conditions: because any default demographic profile is held constant across both conditions for a given model, demographic defaulting cannot account for the systematic shift observed when the instrument became recognizable. Demographically anchored personas matched to a specific norm sample would be needed to make any absolute comparison interpretable. Finally, while the methods aimed to replicate a real-world "evaluation context," the actual dynamics of user interactions with LLMs in clinical or research environments may involve more nuanced cues and pressures.

This study did not measure the effects of item order and model temperature, although previous studies have assessed these variables and established that they have negligible effects [4]. Similarly, acquiescence bias (ie, being more likely to agree with questions) has been shown to be present to some degree in LLM responses [23,24]. However, reverse-coding items has previously been shown to reduce bias by only around 50% [4]. It should also be noted that trial-to-trial variance in model responses was small (total score SDs of approximately 1-5 points), so the 100 trials per condition are better regarded as repeated samples from a near-deterministic response distribution than as 100 independent respondents. The *t* tests establish that the 2 conditions differ reliably for a given model, but the very large standardized effect sizes partly reflect this low variance and should not be interpreted as they would be for a human sample.

Future Work

Two extensions follow directly from the limitations above. The pipeline used here can be applied without modification to models from other developers; replication with at least 1 non-OpenAI model family, and with current-generation models, is required before any general claim about LLMs can be made. Second, administering the instruments under demographically anchored personas matched to published norm samples would allow absolute scores, and not only the between-condition shift, to be interpreted. Beyond these, the "politeness skew" of LLMs should be probed in domains where its impact is subtler but equally consequential. There is a need for further systematic audits of diversity and representation to establish, for example, whether synthetic patient cases underrepresent minoritized sexual orientations, neurodivergent presentations, or intersectional identities [21]. Similarly, evaluations of LLMs' emotional intelligence should move beyond sentiment classification to measure how convincingly models can simulate ambivalent, conflicted, or socially disapproved emotions (eg, resentment toward medical staff or ambivalence about sobriety). Classic technical benchmarks (factual recall, reasoning chains, and robustness to adversarial prompts) should also be reevaluated under conditions explicitly designed to elicit norm-violating content; preliminary evidence suggests accuracy degrades when truthful answers carry reputational risk [25].

XSL•FO
RenderX

Finally, the "naive" condition involved presenting scale items in isolation. While this successfully blinded the model to the evaluative context, it disrupts the psychometric structure of the tools. It is possible that the absence of surrounding items altered response patterns through mechanisms other than SDB, such as the loss of central tendency often seen in full-scale completion. More fundamentally, because the naive and informed conditions differ simultaneously in evaluative transparency, prompt length, and aggregate item sentiment, the present design cannot definitively establish that the observed shift is mediated by the model's inference of an evaluative context rather than by context-length anchoring or sentiment equalization. The interpretation in terms of SDB should therefore be regarded as the most parsimonious account consistent with prior work rather than as a directly demonstrated mechanism. Future work could disentangle these explanations by holding prompt length and sentiment constant while varying only evaluative framing, for example, by embedding scale items within length-matched "filler" items of neutral valence or by explicitly signaling vs concealing the evaluative purpose at fixed context length. A complementary and more direct approach would be to construct application-level test harnesses that task models with the specific job of concern (eg, generating synthetic patient cohorts or simulating survey respondents for biomedical research) and then measure SDB in those generated data against ground truth distributions. Demonstrating the bias within the end-use task itself, rather than inferring it from screening questionnaire responses, would more directly establish the practical risk to LLM-driven data generation.

Conclusions

In 4 OpenAI models, presenting a validated health screening instrument in a form that allowed it to be recognized shifted responses toward the socially desirable pole in 14 of 16 model-instrument combinations, with the direction and size of the shift depending on the instrument and the model family. These results provide formative evidence for the hypothesis that SDB, previously documented in LLM responses to personality inventories, extends to clinical screening instruments; they do not establish it as a general property of LLMs, and the mechanism remains to be isolated. Two practical safeguards follow regardless of mechanism: prompt-level bias audits before an LLM is used as a screening or data generation tool in health settings and explicit reporting of prompt framing in studies that use LLM-generated health data, as framing alone systematically shifted responses in this study.

SDB in LLMs has been documented in personality and survey settings; these results extend that observation to validated health screening instruments for one vendor's models, highlighting its potential impact on research, clinical practice, and medical education. Whether the propensity persists, intensifies, or resolves as models evolve and whether it is shared across developers are empirical questions that the pipeline described here is designed to answer. Understanding (and ultimately mitigating) this bias is a prerequisite for the valid use of LLMs as simulated patients or synthetic respondents in health care, psychology, and related fields.

Acknowledgments

The author wishes to thank his colleagues for their assistance with proofreading this manuscript and providing advice on the statistical methods.

Funding

The author declared that no financial support was received for this work.

Data Availability

The complete dataset for this study is available in the University of St Andrews data repository [26].

Generative AI Use

Generative AI was used in the preparation of this manuscript: Claude (version Fable 5; Anthropic PBC) was used to proofread the original submission and to assist with language editing and restructuring of the revised manuscript in response to reviewer comments.

Authors' Contributions

This is a single-author study. Therefore, AO undertook conceptualization, methodology, software, formal analysis, investigation, data curation, visualization, project administration, writing—original draft, and writing—review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Prompts data.
[PDF File (Adobe PDF File), 138 KB-Multimedia Appendix 1]

XSL•FO
RenderX

References

1. Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, et al. Emergent abilities of large language models. arXiv. Preprint posted online on June 15, 2022. [FREE Full text] [doi: 10.48550/arXiv.2206.07682]
2. Beins BC. Social desirability bias. In: Keith KD, editor. The Encyclopedia of Cross-Cultural Psychology. Malden, MA. Wiley-Blackwell; Oct 17, 2013:1203-1205.
3. Grimm P. Social desirability bias. Part 2: marketing research. In: Sheth J, Malhotra N, editors. Wiley International Encyclopedia of Marketing. Hoboken, NJ. John Wiley & Sons; 2010.
4. Salecha A, Ireland ME, Subrahmanya S, Sedoc J, Ungar LH, Eichstaedt JC. Large language models display human-like social desirability biases in Big Five personality surveys. PNAS Nexus. Dec 17, 2024;3(12):pgae533. [FREE Full text] [doi: 10.1093/pnasnexus/pgae533] [Medline: 39691446]
5. Neeley SM, Cronley ML. When research participants don't tell it like it is: pinpointing the effects of social desirability bias using self vs. indirect-questioning. Adv Consum Res. 2004;31(1):432-434. [FREE Full text] [doi: 10.4135/9781412950589.n932]
6. Holtgraves T. Social desirability and self-reports: testing models of socially desirable responding. Pers Soc Psychol Bull. Feb 2004;30(2):161-172. [doi: 10.1177/0146167203259930] [Medline: 15030631]
7. Nederhof AJ. Methods of coping with social desirability bias: a review. Eur J Soc Psychol. 1985;15(3):263-280. [doi: 10.1002/ejsp.2420150303]
8. Widmar NJ, Byrd ES, Dominick SR, Wolf CA, Acharya L. Social desirability bias in reporting of holiday season healthfulness. Prev Med Rep. Jun 29, 2016;4:270-276. [FREE Full text] [doi: 10.1016/j.pmedr.2016.06.017] [Medline: 27453811]
9. Burke MA, Carman KG. You can be too thin (but not too tall): social desirability bias in self-reports of weight and height. Econ Hum Biol. Nov 2017;27(Pt A):198-222. [doi: 10.1016/j.ehb.2017.06.002] [Medline: 28768226]
10. Latkin CA, Edwards C, Davey-Rothwell MA, Tobin KE. The relationship between social desirability bias and self-reports of health, substance use, and social network factors among urban substance users in Baltimore, Maryland. Addict Behav. Oct 2017;73:133-136. [FREE Full text] [doi: 10.1016/j.addbeh.2017.05.005] [Medline: 28511097]
11. Stuart GS, Grimes DA. Social desirability bias in family planning studies: a neglected problem. Contraception. Aug 2009;80(2):108-112. [doi: 10.1016/j.contraception.2009.02.009] [Medline: 19631784]
12. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing Generalized Anxiety Disorder: the GAD-7. Arch Intern Med. May 22, 2006;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
13. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. Sep 2001;16(9):606-613. [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
14. Saunders JB, Aasland OG, Babor TF, de la Fuente JR, Grant M. Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption--II. Addiction. Jun 1993;88(6):791-804. [doi: 10.1111/j.1360-0443.1993.tb02093.x] [Medline: 8329970]
15. Wilson DM, Nielsen E, Ciliska D. Lifestyle assessment: testing the FANTASTIC instrument. Can Fam Physician. Sep 1984;30:1863-1866. [FREE Full text]
16. Löwe B, Decker O, Müller S, Brähler E, Schellberg D, Herzog W, et al. Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. Med Care. Mar 2008;46(3):266-274. [doi: 10.1097/MLR.0b013e318160d093] [Medline: 18388841]
17. Kocalevent RD, Hinz A, Brähler E. Standardization of the depression screener Patient Health Questionnaire (PHQ-9) in the general population. Gen Hosp Psychiatry. Sep 2013;35(5):551-555. [doi: 10.1016/j.genhosppsych.2013.04.006] [Medline: 23664569]
18. Sanchez-Roige S, Fontanillas P, Elson SL, 23andMe Research Team, Gray JC, de Wit H, et al. Genome-wide association study of Alcohol Use Disorder Identification Test (AUDIT) scores in 20 328 research participants of European ancestry. Addict Biol. Jan 2019;24(1):121-131. [FREE Full text] [doi: 10.1111/adb.12574] [Medline: 29058377]
19. Simpson R, Albert W, Wilson DM, Ciliska D, Evans CE. Lifestyle assessment: part 4. The Halton Health Promotion Survey. Can Fam Physician. Oct 1984;30:2147-2155. [FREE Full text]
20. Lee S, Yang KQ, Peng TQ, Heo R, Liu H. Exploring social desirability response bias in large language models: evidence from GPT-4 simulations. arXiv. Preprint posted online on October 20, 2024. [FREE Full text] [doi: 10.48550/arXiv.2410.15442]
21. O'Malley A, Veenhuizen M, Ahmed A. Ensuring appropriate representation in artificial intelligence-generated medical imagery: protocol for a methodological approach to address skin tone bias. JMIR AI. Nov 27, 2024;3:e58275. [FREE Full text] [doi: 10.2196/58275] [Medline: 39602221]
22. Veenhuizen M, O'Malley A. Demographic biases in AI-generated simulated patient cohorts: a comparative analysis against census benchmarks. Adv Simul (Lond). Nov 18, 2025;10(1):58. [FREE Full text] [doi: 10.1186/s41077-025-00385-9] [Medline: 41254805]
23. Mei Q, Xie Y, Yuan W, Jackson MO. A Turing test of whether AI chatbots are behaviorally similar to humans. Proc Natl Acad Sci U S A. Feb 27, 2024;121(9):e2313925121. [FREE Full text] [doi: 10.1073/pnas.2313925121] [Medline: 38386710]

XSL•FO
RenderX

24. Dentella V, Günther F, Leivada E. Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias. Proc Natl Acad Sci U S A. Dec 19, 2023;120(51):e2309583120. [FREE Full text] [doi: 10.1073/pnas.2309583120] [Medline: 38091290]
25. Schmidt HG, Rotgans JI, Mamede S. Bias sensitivity in diagnostic decision-making: comparing ChatGPT with residents. J Gen Intern Med. Mar 2025;40(4):790-795. [doi: 10.1007/s11606-024-09177-9] [Medline: 39511117]
26. O'Malley A. Social desirability bias in large language models completing health screening questionnaires: experimental study of naïve and informed prompts across four OpenAI models. JMIR Form Res. Preprint posted online on October 13, 2026. [FREE Full text] [doi: 10.2196/92057]

Abbreviations

AUDIT: Alcohol Use Disorders Identification Test
GAD-7: Generalized Anxiety Disorder-7
LLM: large language model
PHQ-9: Patient Health Questionnaire-9
SDB: social desirability bias

Edited by I Steenstra; submitted 23.Jan.2026; peer-reviewed by AL Glowinski, R Huang, K Sarma; comments to author 15.May.2026; revised version received 13.Oct.2026; accepted 13.Oct.2026; published 03.Nov.2026

Please cite as:
O'Malley A
Social Desirability Bias in Health Screening Questionnaire Responses Generated by OpenAI Models: Quantitative Experimental Study
JMIR Form Res 2026;10:e92057
URL: https://formative.jmir.org/2026/1/e92057
doi: 10.2196/92057
PMID:

©Andrew O'Malley. Originally published in JMIR Formative Research (https://formative.jmir.org), 03.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX