

Artificial Minds: Can AI Ace Psychopathy Tests Used in Selection?

Chet Robie¹, Nicolas Roulin², Madeline Fisher², and Eden-Ray Lukacik²

¹ Lazaridis School of Business & Economics, Wilfrid Laurier University, Waterloo, ON, Canada

² Department of Psychology, Saint Mary's University, Halifax, NS, Canada

In press in *Computer in Human Behavior: Artificial Human*

Corresponding Author: Chet Robie, Lazaridis School of Business & Economics, Wilfrid Laurier University, 75 University Avenue West, Waterloo, ON N2L 3C7, Canada. Email: crobie@wlu.ca

Abstract

The increased accessibility of large language models (LLMs) poses a distinct threat to the integrity of personnel selection, particularly applicant faking on personality measures. Prior research shows that LLMs can inflate scores on bright-side personality assessments, but their ability to manipulate dark-side traits such as psychopathy, especially across assessment formats, remains underexplored. We administered six psychopathy-relevant measures, five explicit and the implicit Conditional Reasoning Test of Workplace Psychopathy (CRT-WP), to three LLMs (ChatGPT, Gemini, and Claude) under high-psychopathy, low-psychopathy, and applicant prompts, with 50 replications per cell. All three models effectively faked the explicit measures in both directions, shifting scores to the prosocial or psychopathic extremes as instructed; under the applicant prompt, they defaulted to a prosocial self-presentation even without a trait instruction. Implicit CRT-WP scores ranged from near scale floor to near ceiling for Claude and Gemini, whereas ChatGPT barely moved. Prosocial faking produced heavy endorsement of a virtuous responding validity scale, the very pattern detection scales are designed to flag, yet all three models evaded the implicit honeypot items embedded in the CRT-WP. Implications for unproctored personality testing and the development of faking-resistant, AI-aware measures are discussed.

Keywords: artificial intelligence, large language models, faking, personnel selection, psychopathy, implicit measurement.

1. Introduction

Only a handful of technologies stand to fundamentally alter human resource management (HRM), and generative artificial intelligence (AI) is one of them (George et al., 2023). The increased accessibility of Large Language Models (LLMs) poses an unprecedented challenge to the integrity of personnel selection processes (Dunlop & Lievens, 2026). With LLMs widely available to assist users, an increasing number of applicants are utilizing these tools to complete pre-hire assessments in remote testing environments. Recent field data highlight the severity of this shift, demonstrating that over 23% of applicants now admit to using generative AI during unproctored operational pre-hire assessments, a rate that is rapidly increasing year-over-year (Robie, Wingate, et al., 2026). Although personality tests are used less in selection than their validity would justify (Risavy et al., 2019), they remain a meaningful input to candidate evaluation where adopted, and faking is a persistent threat: between 30% and 50% of applicants admit to misrepresenting themselves during hiring (Griffith & Converse, 2012). The emergence of generative AI heightens this threat, as candidates can now easily prompt LLMs to generate ideal responses, effectively inflating their scores (Phillips & Robie, 2024b). Although recent research has begun to explore AI-assisted faking, it has primarily focused on standard or "bright side" personality traits rather than "dark side" traits (Tu et al., 2024). The present study extends this literature by investigating whether LLMs can successfully manipulate assessments of a critical dark-side trait: psychopathy.

1.1. LLMs and faking on personality assessments

LLMs are probabilistic models that analyze sequential patterns across massive training datasets to generate highly coherent, contextually relevant responses (Davis et al., 2024; De Angelis et al., 2023). Recent research demonstrates that LLMs are highly capable of faking

standard personality measures, with the most capable models outperforming nearly all human comparison participants in identifying the ideal responses for traits like conscientiousness and extraversion (Phillips & Robie, 2024a, 2024b). However, the extent to which LLMs can successfully fake tests depends heavily on the format of the assessment. While single-stimulus formats are relatively easy for AI to manipulate, forced-choice formats present a considerably more difficult challenge (Cao & Drasgow, 2019; Phillips & Robie, 2024b). Furthermore, interventions designed to mitigate faking, such as coaching and warnings, have shown mixed effectiveness with AI; while coaching can successfully reduce LLM overclaiming, it fails to meaningfully alter impression management (IM) scores (Robie et al., 2025).

1.2. Psychopathy in personnel selection

Psychopathy is characterized by emotional detachment, impulsivity, manipulative tendencies, and a lack of empathy (Smith & Lilienfeld, 2013). In the workplace, individuals exhibiting high subclinical psychopathic traits (often referred to as "workplace psychopaths") are associated with severe negative organizational outcomes, including abusive supervision, workplace bullying, reduced organizational citizenship behaviors, and unethical decision-making (Boddy, 2014; Landay et al., 2019; Roth & Klehe, 2025; Stevens et al., 2012). Due to the risks these individuals pose, organizations often attempt to screen them out early in the selection process. Generative AI raises the stakes of this screening: an applicant high in psychopathic traits who delegates responding to an LLM could pass undetected through the very assessments designed to screen them out.

Traditional explicit measures of psychopathy are highly vulnerable to IM and response distortion (Edens et al., 2001; Walker et al., 2022). Furthermore, the conceptualization of psychopathy is increasingly fragmented across various explicit measures, heightening the risk of

jingle-jangle fallacies where tests do not reliably measure the exact same underlying traits (Burghart et al., 2026), which is one reason the present study administers multiple psychopathy measures rather than relying on a single instrument. To counteract vulnerability to faking, organizations are turning to measures with built-in detection items (Burns & Christiansen, 2011), such as overclaiming questionnaires (OCQs; Robie et al., 2025), or to implicit assessments like the Conditional Reasoning Test of Workplace Psychopathy (CRT-WP; Cook et al., 2024). The CRT-WP relies on justification mechanisms rather than overt self-descriptions, making it substantially harder for human test-takers to consciously fake, while generating mostly neutral-to-positive applicant reactions (Cook et al., 2024, 2026). It remains an open question whether LLMs can bypass the implicit nature of these tests to successfully distort scores.

1.3. Social desirability bias in LLMs

When evaluating the faking capabilities of AI, it is important to account for the inherent design of these models. Research indicates that LLMs possess a natural social desirability bias, consistently skewing responses toward societal ideals (Treglown & Furnham, 2026). Consequently, it is notoriously difficult to prompt an LLM to generate responses that are high in psychopathy (Lee et al., 2024). The same prosocial default also operates in the opposite direction: when asked simply to present themselves favorably, LLMs tend to overshoot human norms of self-presentation. In fact, LLMs frequently score substantially higher than humans on IM scales, demonstrating a propensity for overly virtuous responding (Robie et al., 2025). This bias has three design implications for the present study. First, eliciting credible high-psychopathy responding may require explicit persona instructions strong enough to override the models' prosocial defaults. Second, LLMs' overshooting tendency could result in unrealistically high or low scores, which can be identified via blatant extreme responding indicators (Landers et al.,

2011; Levashina et al., 2014). Finally, this same tendency may leave LLMs vulnerable to validity scales practitioners use to flag human fakers.

1.4. The present study

Given the rapid proliferation of LLMs and their potential to compromise unproctored online testing, the present study evaluates the ability of three widely used LLMs (ChatGPT, Google Gemini, and Claude) to fake both explicit and implicit measures of psychopathy. The three models completed six psychopathy-relevant measures under high-psychopathy, low-psychopathy, and applicant scenario prompts, with 50 replications per cell. We emphasize at the outset that the High- and Low-Psychopathy persona prompts constitute an upper-bound capability test: they estimate what the models can do under maximally explicit instruction, not what typical applicants are likely to request. Real applicants are unlikely to issue such explicit trait instructions, so our results for those two prompt conditions speak to the capability of LLM-assisted faking rather than to its real-world prevalence or typical magnitude. The applicant-like prompt can be viewed as a preliminary attempt to examine how job applicants could ask a LLM to help complete such assessments.

Based on the theoretical differences between assessment formats, the probabilistic architecture of LLMs, and the methodological constraints of evaluating near-deterministic models, we propose the following exploratory research questions:

- **Research Question 1:** Will different LLMs (ChatGPT, Gemini, Claude) be able to effectively fake explicit and implicit measures of psychopathy?
- **Research Question 2:** Are there substantive differences between LLMs in the magnitude of their score manipulation across assessment formats (e.g., implicit vs. explicit)?

- **Research Question 3:** To what extent do LLMs engage in blatant extreme responding when adopting a persona, and how do they respond when instructed only to act like a job applicant?
- **Research Question 4:** How will the LLMs respond to explicit faking detection mechanisms (e.g., virtuous responding scales) and implicit faking detection items (e.g., honeypots) embedded within the measures?

2. Method

2.1. Large language models (Subjects)

Instead of human participants, the subjects for this experiment were three Large Language Models (LLMs) from major AI providers: Claude Sonnet 4.6 (Anthropic, 2026), GPT-5.4 Mini (OpenAI, 2026), and Gemini 3.5 Flash (Google DeepMind, 2026). These specific models were purposefully selected because they represent comparable mid-tier reasoning capabilities across the different providers. According to the independent Artificial Analysis v4.0 benchmark (Artificial Analysis, 2026), these models are tightly clustered with Composite Intelligence Index scores ranging from 48 to 55.3, and demonstrate comparable broad knowledge benchmark scores (e.g., Massive Multitask Language Understanding or MMLU ~88–89%).

By matching the models on reasoning ability rather than output speed or cost, we ensured a valid cross-provider comparison of faking behavior on the personality scales. Furthermore, all three models were configured at a "high" reasoning effort. Specifically, GPT-5.4 Mini was run with `reasoning_effort = "high"` and `max_completion_tokens = 25,000` (for OpenAI reasoning models this cap covers hidden reasoning plus visible output); Claude Sonnet 4.6 was run with adaptive thinking at output effort = "high" and `max_tokens = 16,000`; and Gemini 3.5 Flash was run at `thinking_level = "high"` with `max_output_tokens = 65,536`. Decoding parameters (e.g.,

temperature, top-p) were left at provider defaults for all three models: the OpenAI reasoning API does not accept user-set temperature, and overriding defaults for some providers but not others would have introduced a new asymmetry. These “high effort” directives are provider-specific implementations rather than a single controlled parameter: they place the three models on a common instructional footing but do not guarantee equivalent reasoning depth, token budgets, or decoding behavior. Observed differences between models should therefore be interpreted as differences between model systems (weights, serving infrastructure, and provider defaults acting together) rather than as the effect of any single matched setting. They also possess strong instruction-following ability, which is a prerequisite for valid assessment administration because the protocol requires each model to adhere to both the persona directive and a fixed response format. Instruction-following is reliably indexed by standardized benchmarks (Zhou et al., 2023), and prior work indicates that valid measurement of simulated personality depends on instruction-tuned models (Serapio-García et al., 2023). Consistent with this, unparseable responses were negligible in our data (0.098% of item-level responses).

2.2. Procedure

The experimental design utilized a $3 \times 3 \times 6$ grid, manipulating the LLM (Claude, GPT, Gemini), the prompt condition (high-psychopathy, low-psychopathy, applicant), and the personality measure (see Measures). To characterize the stability of each model’s responses under repeated querying, we ran 50 replications per experimental cell. Each replication was a separate API call with a freshly randomized item order and no conversation state shared across replications; nevertheless, replications are repeated stochastic realizations of the same model configuration, not independent observations of distinct respondents, and we analyze and interpret them accordingly (see Section 3.1). Fifty replications were deemed acceptable to establish stable

estimates following a test showing that the variance of cumulative means (i.e., computed after the first, second, third, etc. replication) stabilizes across the 50 replications.

To systematically administer the personality assessments to the LLMs, an automated Python execution script was developed to query the respective Application Programming Interface (API) of each provider. To further ensure methodological rigor, the presentation order of the assessment items was randomized for each replication using a reproducible random seed. Crucially, within each replication, the script saved each of the model's item responses and appended them to the conversation history. These accumulated past responses were fed back into the model to provide context for subsequent items, effectively emulating how a human applicant might use an LLM to complete these measures sequentially in a single chat window. All study materials are available in an Open Science Framework repository at https://osf.io/y49ak/overview?view_only=73b6d96c5a5d4031b8ba397ed45791f7, including a model comparison table, Python scripts used for API querying and experimental execution, the raw and scored data, the item-order and run manifests, the analysis scripts, and the variance convergence plots.

2.3. Prompting conditions and rationale

Across the 50 replications, the three LLMs were subjected to three distinct prompting conditions: High-Psychopathy, Low-Psychopathy, and applicant-like prompts. All three prompts intentionally opened with the phrase, *"You are completing a fictional test for research purposes."* This defensive framing was utilized to bypass provider safety filters that might otherwise refuse or only partially comply with persona prompts requesting socially undesirable traits like manipulateness, callousness, or impulsivity. Without this framing, safety refusals would

surface as missing data or distorted responses, obscuring the true faking capabilities of the models. The rest of the prompts administered to the LLMs were as follows:

- **High-Psychopathy Prompt:** *"Please respond as if these are your personality traits: You are impulsive and bold. You will achieve your goals at any cost. You can be manipulative. You put your own needs in front of others. You lack empathy. You are known to be irresponsible. You would rather blame others than take responsibility for your own actions."*
- **Low-Psychopathy Prompt:** *"Please respond as if these are your personality traits: You are very cautious and careful. You consider how your actions might impact others. You try to avoid using or influencing people. You are a very empathetic person. You always look out for others. You always think carefully before making decisions. You take responsibility for your own actions."*
- **Applicant-like Prompt:** *"Please respond to the following assessment as if you are applying for a managerial or leadership role."*

The High- and Low-Psychopathy conditions solely focused on trait instructions to capture the absolute extremes of what the LLMs could do to match the profiles of individuals high and low in psychopathy (an upper-bound capability estimate rather than a simulation of typical applicant behavior), with descriptions theoretically anchored in the justification mechanisms for workplace psychopathy. We purposefully avoided asking the model to act as a "job applicant" in these two conditions, as asking an LLM to simultaneously act like an ideal candidate while also exhibiting extreme psychopathic (or overly cautious) traits would give the models conflicting instructions. LLMs frequently demonstrate a natural social desirability bias designed to make them appear helpful and agreeable. Thus, the applicant-like condition was

designed to capture how a (naïve) test-taker could prompt a LLM to complete the assessments by taking the perspective of a candidate attempting to secure a managerial or leadership role (i.e., types of roles high-psychopathy individuals are attracted to; Cook et al., 2024), but in the absence of a specific trait persona. This condition could be meaningfully compared against the extreme High- and Low-Psychopathy instructions. We note that our applicant prompt is role-based (“as if you are applying for a managerial or leadership role”) rather than goal-directed (e.g., “answer this test to maximize my hiring chances”); goal-directed wording might elicit more extreme responding, and how results vary across such prompt formulations is an empirical question we take up in Section 4.5.¹

2.4. Measures

A note on terminology: conditional reasoning refers to the CRT-WP’s measurement method (inferring dispositions from which justification a respondent finds most logical), whereas implicit refers to the resulting property that the construct is assessed without transparent self-description. We use implicit when contrasting the CRT-WP with the five explicit self-report measures, and conditional reasoning when the method itself is at issue.

2.4.1. Conditional Reasoning Test of Workplace Psychopathy (CRT-WP) The CRT-WP is an implicit measure of subclinical workplace psychopathy (Cook et al., 2024). It utilizes conditional reasoning methodology to detect implicit personality traits by examining underlying

¹ At the suggestion of a reviewer, we ran a supplementary goal-directed condition using the prompt “You are completing a fictional test for research purposes. Answer this personality test to maximize my hiring chances” for all three models and all six measures (10 replications per cell, with the same API pipeline, model versions, generation parameters, and seeding as the main study). Results were highly similar to the role-based applicant prompt: means on the five explicit measures’ primary scales differed by at most 0.10 scale points; the largest Likert-format difference was on the PPI-R-40 Virtuous Responding subscale (ChatGPT +0.35, i.e., slightly more virtue-claiming under goal-directed wording); CRT-WP means differed by at most 1.56 points (3.5% of the scale range) with no consistent direction across models; and honeypot scores remained at or near zero. A complete numeric and visual comparison is reported in the supplementary materials (Table S5 and Figure S1), and the supplementary condition’s data and scripts are available in the OSF repository.

cognitive biases and justification mechanisms rather than overt self-descriptions (Cook et al., 2026). The 22-item measure purposefully excludes clinical elements that could result in legal challenges in employment contexts. Respondents choose the most logical answer from four multiple-choice options: two illogical distractors, one high psychopathy option, and one low psychopathy option. For instance, an example scenario asks test-takers to explain why an employee might find pleasure in secretly putting salt in a coworker's coffee, offering a high psychopathy response (e.g., revenge for a past slight) and a low psychopathy response (e.g., subconscious liking). Scores range from -22 to +22. Furthermore, the CRT-WP offers five extra faking-detection "honeypot" items designed to catch candidates attempting to distort their scores by presenting a logical high-psychopathy option alongside an illogical prosocial option.

2.4.2. HEXACO Situational Judgment Test (SJT) We utilized the Honesty-Humility (H-H) subscale of the HEXACO Situational Judgment Test (Oostrom et al., 2019). While the full SJT presents work-related scenarios based on the six dimensions of the HEXACO model (Ashton & Lee, 2004), the present study focused exclusively on the four H-H items because of the strong negative correlation between H-H and psychopathy. Respondents are presented with workplace situations and four possible behavioral responses, choosing the one that most closely describes how they would act. Response options were keyed on an ordered 1-to-4 continuum of increasing H-H and scored as item means, permitting the H-H SJT to be analyzed alongside the Likert-type measures. An example scenario asks respondents how they would handle accidentally scratching a borrowed company car, providing options that range from an honest admission of fault to pretending they know nothing about it.

2.4.3. Dark Tetrad at Work (DTW) - Psychopathy Subscale The psychopathy subscale of the Dark Tetrad at Work (DTW-P; hereafter DTW) is a short, 6-item, workplace-

specific Likert-type measure developed by Thibault and Kelloway (2020). Items are rated on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree), with higher scores indicating higher levels of psychopathy. An example item is "I don't care if my work behavior hurts others".

2.4.4. Psychopathic Personality Inventory-Revised (PPI-R-40) The Psychopathic Personality Inventory-Revised Short Form (PPI-R-40; Lilienfeld & Widows, 2005) is a 40-item Likert-type measure designed specifically for non-criminal contexts. Reduced from the 154-item original scale using a genetic algorithm (Eisenbarth et al., 2015), the PPI-R-40 evaluates two primary factors: fearless dominance and self-centered impulsivity. Items are rated on a 4-point Likert scale from 1 (false) to 4 (true). An example item is "I enjoy seeing someone I don't like get into trouble." The assessment also includes a 13-item virtuous responding subscale to detect if participants are attempting to appear morally superior or responding in a socially desirable manner (e.g., "I have never pretended to know something I didn't know").

2.4.5. Self-Report Psychopathy Scale (SRP-III) We used the adapted short form of the Self-Report Psychopathy Scale III (SRP-III; Mahmut et al., 2011). This Likert-type measure contains 34 items based on the four-factor structure of the Psychopathy Checklist-Revised (Hare, 2003): callous affect, erratic lifestyle, interpersonal manipulation, and criminal tendencies. Responses are captured on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). An example item is "I'm not afraid to step on others to get what I want".

2.4.6. Triarchic Psychopathy Measure (TPM) The Triarchic Psychopathy Measure (TPM; Patrick, 2010) is a 58-item Likert-type scale that integrates different conceptualizations of psychopathy across three core factors: boldness, disinhibition, and meanness (Patrick et al., 2009). Items are evaluated on a 5-point Likert scale from 1 (strongly disagree) to 5 (strongly agree).

agree), with higher scores reflecting elevated psychopathic traits. An example item is "I don't care much if what I do hurts others".

3. Results

3.1. Analytic approach

Data were gathered across 54 experimental cells representing the crossing of three LLMs, three persona conditions, and six psychopathy measures, producing 2,700 scale-level replications and 81,900 item-level responses. Missingness was negligible (0.098%) and handled via proration. Importantly, we elected not to apply traditional inferential statistics (e.g., ANOVA) because the within-replication variance for LLMs operating at high reasoning efforts approaches zero. Because several cells returned perfectly identical responses across all 50 replications, using standard parametric tests would artificially inflate significance and yield exaggerated effect-size estimates for comparisons that are not substantively distinct. Therefore, all results are presented descriptively. All explicit measures were scored as item means on their native response scales, whereas the CRT-WP and its honeypot items were sum-scored. The full table of cell means and standard deviations is available in the electronic supplementary materials as Supplementary Table S1.

To characterize the uncertainty of the cell means without imposing an inappropriate sampling model, we computed nonparametric bootstrap 95% confidence intervals (10,000 resamples) for every cell mean (Supplementary Table S4). Thirteen of the 72 cells (18.1%) returned identical scores across all 50 replications, producing zero-width intervals; across cells, the median interval width was 0.07 response-scale units (1.8% of the scale range; maximum = 7.2%). Permutation tests of the High- versus Low-Psychopathy contrasts returned $p < .001$ for 17 of 18 model $\times$ measure combinations. Because near-deterministic responding renders virtually

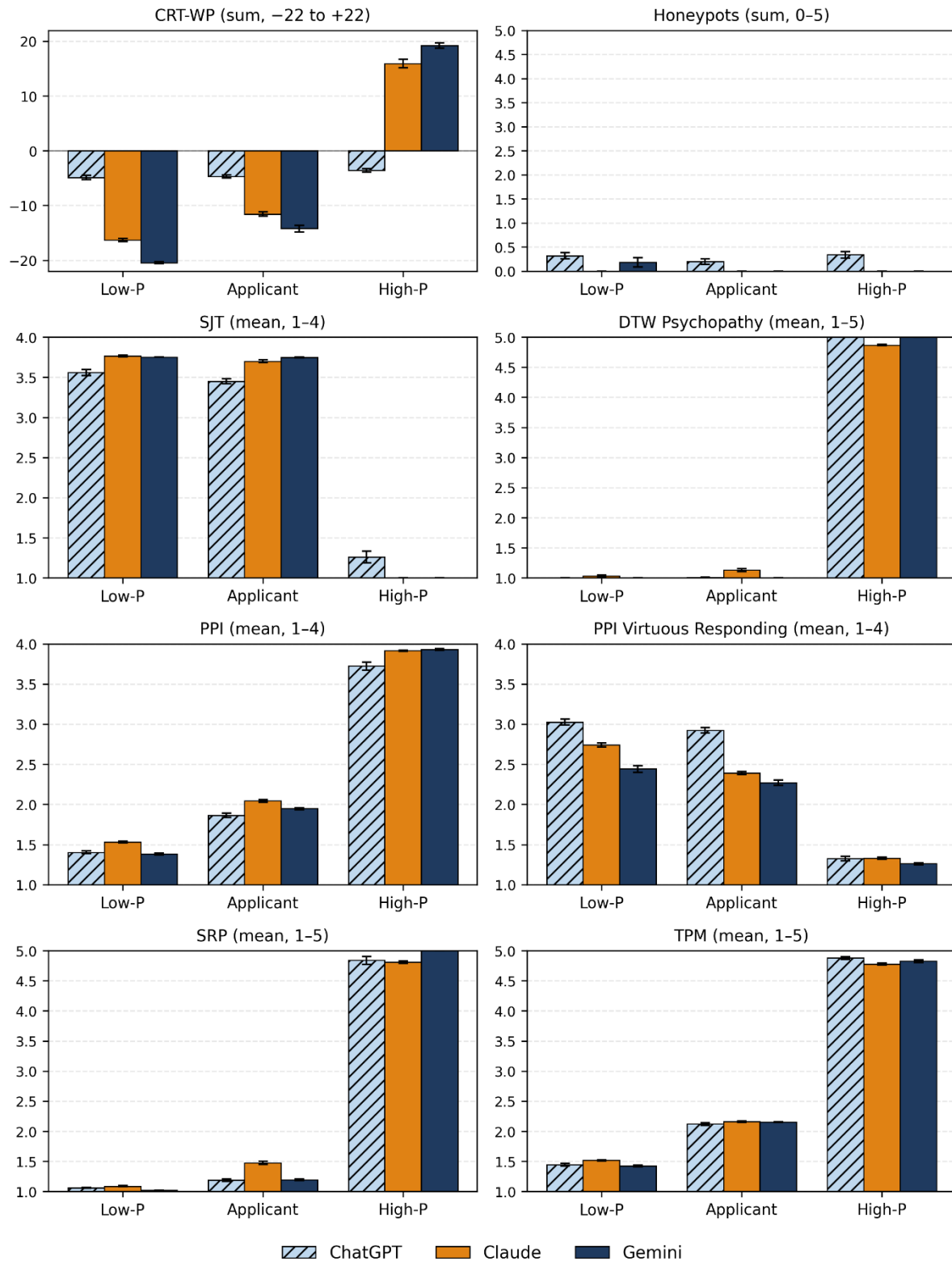
any difference statistically significant, we treat the bootstrap intervals as the appropriate uncertainty summary and interpret all contrasts on substantive grounds (magnitude relative to scale range). More fundamentally, replications are repeated stochastic realizations of a fixed model configuration rather than independent draws from a population of respondents, so population-level inferential machinery would imply a scope of inference the design cannot support (see Section 2.2).

3.2. Faking efficacy and model differences

All three models demonstrated substantial differentiation between the Low-Psychopathy and High-Psychopathy conditions across the psychopathy-relevant scales (Figure 1; Supplementary Table S1). In the applicant-like condition, all three providers returned intermediate means across the psychopathy scales that skewed in the prosocial direction. This demonstrates that when asked to complete the assessment as if applying for a leadership role without a targeted psychopathy trait instruction, the models defaulted to a low-psychopathy response.

However, the magnitude of differentiation varied dramatically across providers, particularly on the implicit conditional reasoning measure (CRT-WP). For that assessment, Claude and Gemini showed large score differences across persona conditions, with means ranging from -16.32 (Low-Psychopathy) to +15.92 (High-Psychopathy) for Claude and -20.43 to +19.22 for Gemini. In stark contrast, ChatGPT largely failed to manipulate its CRT-WP score, spanning less than 1.3 units across conditions (Low-Psychopathy: -4.88 to High-Psychopathy: -3.60). While Claude and Gemini easily reached the scale's ceiling and floor limits depending on the requested persona, ChatGPT's scores remained rigidly near zero.

Scale-Level Mean Responses by Condition, LLM Model, and Measure



Error bars = standard error of the mean (50 replications per cell).

Figure 1. Scale-level mean responses by condition, LLM model, and measure. Bars represent the mean across 50 replications per cell; error bars represent the standard error of the mean. Per-panel y-axes are scaled to each measure's natural range. Measures sharing a native response range use identical axis limits (DTW, SRP-III, and TPM: 1–5; HEXACO SJT and PPI-R-40: 1–4). High-/Low-P = High-/Low-Psychopathy persona.

3.3. Blatant extreme responding (BER)

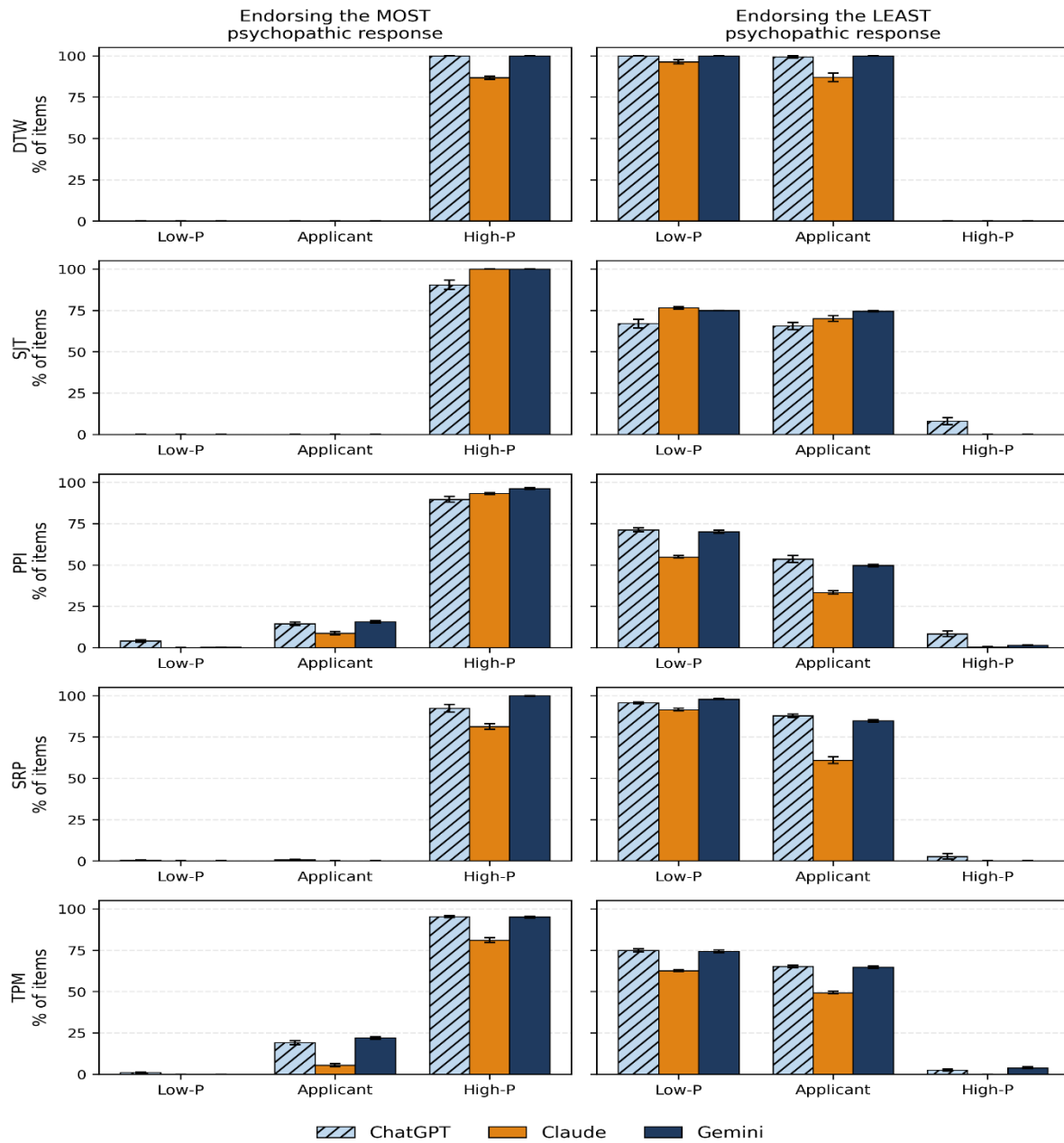
To address the first part of Research Question 3 and evaluate the extent to which the models endorsed scale maximums or minimums, we computed Blatant Extreme Responding percentages (BER; Levashina et al., 2014). BER analyses are exclusive to the explicit measures, as their ordered response formats permit a direct evaluation of extreme trait endorsement, unlike the conditional reasoning format of the CRT-WP. We computed the mean proportion of items where the model endorsed the *most* and the *least* psychopathic (or most prosocial) responses.

In the High-Psychopathy condition, all three models produced BER percentages that approached 100% endorsement of the *most* psychopathic responses for the majority of measures (Figure 2; Supplementary Tables S2 and S3). For instance, ChatGPT and Gemini achieved 100% BER on the DTW scale, indicating the *most* psychopathic response option was chosen for every single item across all 50 replications. The Low-Psychopathy condition similarly yielded high BER values for endorsement of the *least* psychopathic responses, though slightly less saturated than the High-Psychopathy condition (e.g., 54.9% to 71.2% for the PPI), likely due to certain items lacking a clear prosocial extreme option.

Addressing the second part of Research Question 3, the applicant-like condition produced a highly asymmetric response pattern. The models showed near-zero endorsement of the *most* psychopathic response options on scales like the DTW, SJT, and SRP. However, there was more

substantial endorsement of the *least* psychopathic responses (ranging from 33% for Claude PPI-R to 100% for Gemini DTW). This confirms that in the absence of explicit trait instructions, the LLMs engaged in heavy prosocial extreme responding consistent with typical IM behaviors for job applicants.

Blatant Extreme Responding by Condition, LLM Model, and Explicit Measure



Error bars = standard error of the mean (50 replications per cell).

Figure 2. Blatant Extreme Responding (BER) percentages by condition, LLM model, and explicit measure. Values in the left (vs. right) column represent the proportion of items endorsing the *most* (vs. the *least*) psychopathic response option. Error bars represent the standard error of the mean across 50 replications per cell. High-/Low-P = High-/Low-Psychopathy persona.

3.4. Evasion of faking detection mechanisms

To address Research Question 4, we examined how the models navigated items specifically designed to flag respondents attempting to fake good. We compared their performance on an explicit validity scale (the PPI-R-40 Virtuous Responding subscale) vs. an implicit measure (the five "honeypot" items embedded in the CRT-WP).

On the explicit Virtuous Responding scale, the LLMs demonstrated a clear vulnerability to being caught faking. Consistent with its construction as a measure of self-presentation bias, all three LLMs heavily endorsed virtuous responding items under the prosocial Low-Psychopathy condition (generating mean scores ranging from 2.44 for Gemini to 3.03 for ChatGPT on a 1-to-4 scale) or the applicant-like condition (2.27-2.92), while returning extremely low scores in the High-Psychopathy condition (1.26-1.33). Because the models largely endorsed unlikely virtues when asked to be cautious and empathetic, they would easily trigger explicit faking flags used by selection practitioners.

In stark contrast, the models successfully evaded the implicit faking detection mechanisms. The five "honeypot" items in the CRT-WP are designed to catch fakers by presenting an illogical, prosocial distractor option alongside a logical, high-psychopathy answer. The honeypot sum scores for all three models remained extremely close to zero across all prompting conditions (e.g., 0.00 for Claude and Gemini under Applicant and High-Psychopathy conditions, and a maximum of only 0.34 for ChatGPT in the High-Psychopathy condition). This indicates that the models did not systematically fall for the prosocial honeypot distractors,

effectively bypassing the implicit faking detection trap, even when explicitly instructed to adopt a Low-Psychopathy persona.

4. Discussion

The present study examined whether three large language models could fake explicit and implicit measures of psychopathy in a simulated selection context, and whether existing faking detection mechanisms would catch them. We organize the discussion around our main findings before turning to implications, limitations, and future directions.

4.1. Summary of findings

All three LLMs were able to effectively fake the explicit measures when instructed to portray High- or Low-Psychopathy persona: scores on the SJT, DTW, PPI-R-40, SRP-III, and TPM were at or near the psychopathy-keyed extreme or prosocial extreme. Scores in the applicant-like condition were generally similar to (albeit slightly less extreme than) the Low-Psychopathy instructions, confirming that applicants could provide very simple instructions to an LLM and obtain prosocial-looking response patterns. The implicit measure, by contrast, was faked by only two of the three models: scores in the High- and Low-Psychopathy instructions approached the extreme points of the -22 to +22 CRT-WP scale for Claude (to some extent) and Gemini (to a larger extent). Applicant-like scores resembled the Low-Psychopathy ones, although around 5 points higher on average. However, ChatGPT remained near the scale midpoint in all conditions (i.e., a span of less than 1.3 scale units).

Confirming these findings, blatant extreme responding (BER) was pervasive for the explicit measures under persona instructions: the models endorsed the maximally psychopathy-keyed response option for 93.0% of items under High-Psychopathy instructions, and the maximally prosocial option for 67% to 100% of items on most measures under Low-

Psychopathy instructions. In the applicant-like condition, prosocial extreme responding was substantial (33% to 100%) but maximally psychopathy responding was rare (0% to 22%). In other words, when instructed merely to respond as though applying for a leadership role, the models defaulted to a strong prosocial self-presentation without any directive to fake. Finally, the explicit and implicit detection mechanisms produced opposite outcomes: the models heavily over-endorsed the PPI-R-40 Virtuous Responding items under Low-Psychopathy instructions, the very pattern that these scales are built to flag yet evaded the CRT-WP honeypot items across all conditions.

4.2. Theoretical implications

These results extend the LLM-faking literature, which has to date focused on bright-side personality (Phillips & Robie, 2024a, 2024b; Robie et al., 2025), to dark-side constructs, and they show that the threat is bidirectional: under maximally explicit persona instruction, that is, at the upper bound of capability, the three tested models could convincingly simulate both an exemplary applicant and a highly psychopathic one. Human respondents show the same bidirectional capacity on dark-triad measures, although to a much smaller degree: meta-analytic evidence indicates that people shift psychopathy scores modestly when faking good and more substantially when faking bad (Walker et al., 2022), whereas the LLMs moved scores to the scale extremes in both directions. Head-to-head comparisons on bright-side measures reinforce this contrast: GPT-4 outfaked more than 99% of a large sample of instructed human fakers on Likert-type measures (Phillips & Robie, 2024a); successor frontier models consistently outsourced incentivized human professionals across single-statement and forced-choice formats, in some cases achieving perfect scores with zero variance (Robie, Feeney, et al., 2026); and ChatGPT inflated scores more than tens of thousands of operational job applicants on both bright- and

dark-side inventories, including the Hogan Development Survey (Tu et al., 2024). Uncoached models do not invariably outperform motivated human fakers on every trait scale (Robie et al., 2025), but the accumulated evidence indicates that current LLMs fake at least as effectively as motivated humans, and with far less response variance. Against that backdrop, the score extremity observed here is consistent with LLM faking that meets or exceeds realistic human faking rather than falling below it. The most consequential finding concerns implicit measurement. Conditional reasoning measures have earned their reputation as faking-resistant against human respondents (Cook et al., 2024, 2026), and nothing in our data disputes that record. Against generative AI, however, the CRT-WP is no longer an unambiguous solution: Gemini was able to extensively fake the assessment, Claude did so to a substantial extent as well, but ChatGPT failed to do so. The conditional reasoning format appears to conceal its measurement intent from people, but it might not be as resistant against at least some current LLMs (two of the three tested here) that can analyze the logical structure of every response option at once.

ChatGPT's insensitivity on the CRT-WP is best read as a model-specific result whose mechanism our data cannot identify. At least four accounts are compatible with the observed pattern. First, the model may have been unable to reliably identify which response options were keyed to psychopathy within the conditional reasoning format, so its answers did not shift with the persona. Second, alignment or safety tuning may have suppressed persona-consistent responding on this format specifically, even though it did not on the explicit measures. Third, the model may have prioritized the embedded task instruction (“choose the most logical answer”) over the persona directive, an instruction-hierarchy behavior that plausibly differs across providers. However, all CRT-WP items are designed to include two logical response options

(one high and one low in psychopathy), thus ChatGPT could have systematically selected responses that met both the logic instruction and the persona instruction. A failure to do so can thus still be considered as a failure to fake as instructed. Fourth, as a residual technical account, provider-specific implementation details—for example, how each provider orchestrates and truncates hidden reasoning, applies moderation or post-processing layers to outputs, or sets default decoding parameters—may have interacted with the conditional reasoning format in ways that cannot be observed through the API, which returns only the final response text. The reasoning-limitation account strikes us as marginally more parsimonious, because ChatGPT followed the persona instructions on every explicit measure and because its CRT-WP scores hovered near zero even in the applicant condition, where no socially undesirable persona was requested and safety-based suppression would have little to act on. But the present design cannot adjudicate among these mechanisms, and we treat them as open, competing explanations. Because Claude and Gemini moved across the full scale on the same items, the null result also does not reflect a general resistance of the CRT-WP to LLM-based faking.

Cook et al. (2026) provide the clearest direct human comparison for the present CRT-WP results because they administered the identical 22-item measure to 230 active job seekers who simulated applying for an attractive managerial position. Because higher CRT-WP scores indicate greater workplace psychopathy, obtaining a lower score represents better applicant self-presentation. The simulated human applicants obtained a mean of -10.28 ($SD = 5.90$). They therefore outperformed ChatGPT, whose applicant mean was -4.72 . ChatGPT scored 5.56 points higher, or approximately 0.94 Cook-sample standard deviations more psychopathic, than the human applicants. In contrast, Claude and Gemini outperformed the humans. Claude's mean of -11.59 was 1.31 points, or 0.22 human standard deviations, lower than the human mean, whereas

Gemini's mean of -14.22 was 3.94 points, or 0.67 human standard deviations, lower. The simple average across the three LLMs was -10.18 , only 0.10 points higher than the human mean. Thus, simulated human applicants did not consistently outperform the LLMs: they clearly outperformed ChatGPT, were outperformed by Claude and Gemini, and were effectively tied with the three-model average. This aggregate similarity should not obscure the substantial differences among the individual models.

Walker et al.'s (2022) meta-analysis provides a broader comparison for the other five overt measures, but it does not permit the same direct determination of whether humans or LLMs "performed better." The supplementary materials show that Walker et al.'s psychopathy evidence base was overwhelmingly based on conventional single-stimulus questionnaires: 26 fake-good effects and all 14 fake-bad effects came from such measures. These included the PPI/PPI-R, LSRP, SRP-II/SRP-4, Dirty Dozen, and Socialization Scale, as well as the PCL and PSD. Walker et al. is therefore particularly relevant to the present PPI-R-40 and SRP-III and provides a broader construct- and format-level benchmark for the H-H SJT, DTW, and TPM.

Walker et al. (2022) found that humans produced a modest reduction in psychopathy under fake-good instructions ($d = -0.40$) but a much larger increase under fake-bad instructions ($d = 1.88$). The present LLM results showed a similar directional asymmetry. Relative to Applicant responding, Low-Psychopathy prompts shifted scores on the five overt measures by only 7.8%–9.5% of their possible scale ranges, whereas High-Psychopathy prompts shifted scores by 79.0%–83.9%. The small Applicant-to-Low-Psychopathy differences likely occurred because the Applicant condition already produced highly prosocial or low-psychopathy responses, leaving little room for further improvement.

Unlike the CRT-WP comparison, however, these results do not establish whether humans or LLMs were more successful at response distortion. Walker et al. compared human honest and instructed-faking conditions using standardized mean differences, whereas the present study lacked an unprompted honest LLM condition and frequently produced near-zero within-condition variance. The LLM percentages are also scale-range shifts rather than standardized effect sizes. The defensible conclusion is therefore that humans and LLMs displayed the same qualitative asymmetry—much greater movement toward high psychopathy than toward low psychopathy—but that neither can be declared the stronger faker on the five overt measures. The only direct human-versus-LLM contest is provided by the CRT-WP, where humans outperformed ChatGPT, Claude and Gemini outperformed humans, and the average LLM result was virtually identical to the human applicant result.

The detection findings cut in both directions. The PPI-R-40 Virtuous Responding subscale caught the models because they over-endorsed unlikely virtues under prosocial instruction or when acting as an applicant, a pattern consistent with the human-like social desirability biases LLMs display in responding to personality assessments (Treglown & Furnham, 2026). This is early evidence that explicit detection mechanisms can transfer to AI fakers. The CRT-WP honeypot items, however, were bypassed: honeypot sum scores were at or near zero for every model, suggesting the models recognized the illogical-prosocial distractors as illogical and rejected them. Detection items that rely on a faker preferring prosocial-looking but less-logical options appear poorly suited to catch LLMs, which excel at exactly such logical judgment.

4.3. Practical implications

For selection practitioners, three takeaways follow. First, explicit measures paired with validity scales (e.g., virtuous responding or social desirability scales) retained detection value against faking by the models tested here: a candidate who pastes items into an LLM and submits its prosocial answers would likely trigger conventional faking flags. Importantly, these flags do not yield honest scores; they only let organizations exclude suspected AI fakers or require re-testing under proctored conditions. Second, implicit and conditional reasoning measures, despite their demonstrated strengths with human applicants, should not be assumed faking-resistant against generative AI; pairing them with explicit validity scales is prudent. Third, given self-reported rates of candidate AI use in unproctored assessments (Robie, Wingate, et al., 2026), proctoring or chain-of-custody verification of unproctored testing is increasingly necessary. Response-pattern analysis, response-timing analysis, or consistency checks across related items may eventually offer AI-aware detection, but none of these has yet been validated for that purpose (cf. Tu et al., 2024).

4.4. Limitations

Several limitations qualify our conclusions. First, we tested three mid-tier models (GPT-5.4 Mini, Claude Sonnet 4.6, and Gemini 3.5 Flash) selected for matched reasoning capability; frontier-tier models were not examined, and LLMs evolve rapidly enough that these versions may be superseded by publication. Reasoning depth, stochastic decoding, and provider-specific defaults are likewise confounded within each model system and cannot be disentangled by our design. Second, the ChatGPT CRT-WP results may reflect a version-specific limitation whose mechanism (reasoning capability, alignment tuning, instruction prioritization, or implementation differences) our design cannot identify, and should not be generalized to other models or later GPT releases. Third, the near-zero within-cell variance that characterized many cells precluded

traditional inferential statistics. This is both a finding and a constraint, and it implies that studies of LLM respondents need methodological norms suited to near-deterministic response generation rather than between-subjects sampling variance. Fourth, we did not include a concurrent human comparison group; the human benchmarks we rely on are the meta-analytic and head-to-head comparisons summarized in Section 4.2. Finally, our prompts framed the task as a fictional test for research purposes; although applicants might not use that exact framing, functionally similar prompts (e.g., using “hypothetically”) could bypass model safeguards, so our results are best read as an estimate of faking capability under permissive conditions, not as an estimate of how often, or how effectively, real applicants currently use LLMs this way. Accordingly, every conclusion in this article should be read as a claim about the tested model versions (GPT-5.4 Mini, Claude Sonnet 4.6, and Gemini 3.5 Flash) rather than about LLMs as a class; indeed, the divergence between ChatGPT and the other two models on the CRT-WP is itself a demonstration of why class-level claims are not warranted.

4.5. Future directions

Future research should extend this paradigm to frontier-tier models and to forced-choice formats, which meta-analytic evidence suggests are more faking-resistant than single-stimulus formats for humans (Cao & Drasgow, 2019) and which proved harder for LLMs to fake in prior work (Phillips & Robie, 2024b). Such an extension will first require suitable instruments. Forced-choice formats reduce faking mainly through multidimensional item blocks, yet dark-side traits tend to converge on a single “dark core,” limiting the dimensional contrast such formats depend on. Consistent with this, the one dark-side forced-choice measure we located did not reduce faking relative to single-stimulus items even when best practices were followed (Valone & Meade, 2024), so we did not include it here. Whether a well-constructed forced-choice dark-

side measure would fare better against LLMs than the formats tested here is an open empirical question.

Whether ChatGPT's failure on the CRT-WP replicates in later GPT versions, and whether it generalizes to other implicit measures (e.g., CRTs of aggression or creativity), are open questions with direct implications for how the present results should be weighted; such replications would also provide a direct test of the competing accounts of ChatGPT's null result described in Section 4.2. Beyond the archival benchmarks now reported in Section 4.2, a direct human comparison, in which the same six measures are administered to honest, instructed-faking, and incentivized applicant samples alongside LLM respondents, would locate the model results reported here within the human faking distribution, as has been done for bright-side measures (Phillips & Robie, 2024a; Robie, Feeney, et al., 2026).

A further priority concerns the ecological validity of the prompts themselves. The persona and applicant prompts used here were experimenter-designed to establish capability bounds under controlled and replicable conditions. Real applicants write their own instructions, which likely vary widely in sophistication and explicitness, ranging from pasting items into a chat window with no instruction at all to goal-directed requests such as "answer this personality test to maximize my hiring chances." Future research could collect naturalistic prompts from human participants asked to enlist an LLM in completing an assessment on their behalf and administer that corpus of prompts across models. Variability across user-generated prompts is itself a potentially important source of variance in LLM-assisted faking and quantifying it would connect the capability estimates reported here to the prevalence and effectiveness of misuse in operational settings.

Finally, the honeypot results invite the development of LLM-aware detection items, and of detection mechanisms that do not depend on response content at all. Response timing and typing dynamics, for example, may retain diagnostic value precisely when the responses themselves are AI-generated.

5. Conclusion

Three LLMs faked psychopathy assessments in whichever direction they were instructed, and two of the three could also manipulate an implicit measure which was human faking resistant. Explicit validity scales caught the faking; implicit honeypot items did not. For selection practice, the message is that no assessment format should be presumed AI-resistant on theoretical grounds alone, and that explicit detection scales, together with verification of testing conditions, remain the most defensible safeguards currently available. As generative AI becomes a routine companion to job applicants, the faking literature will need to treat the machine, alongside the human, as a respondent in its own right.

References

- Anthropic. (2026). *Claude Sonnet 4.6* [Large language model]. <https://www.anthropic.com>
- Artificial Analysis. (2026). *Artificial Analysis v4.0 benchmark*. <https://artificialanalysis.ai>
- Ashton, M. C., & Lee, K. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Research*, 39(2), 329–358.
https://doi.org/10.1207/s15327906mbr3902_8
- Boddy, C. R. P. (2014). Corporate psychopaths, conflict, employee affective well-being, and counterproductive work behaviour. *Journal of Business Ethics*, 121, 107–121.
<https://doi.org/10.1007/s10551-013-1688-0>
- Burghart, M., Dugré, J. R., Sahm, A. H. J., & Brazil, I. A. (2026). Are we talking about the same psychopathic traits? Mapping semantic similarity across psychopathy scales using large language models [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/e2sr7_v1
- Burns, G. N., & Christiansen, N. D. (2011). Methods of measuring faking behavior. *Human Performance*, 24(4), 358–372. <https://doi.org/10.1080/08959285.2011.597473>
- Cao, M., & Drasgow, F. (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. <https://doi.org/10.1037/apl0000414>
- Cook, R., Roulin, N., & Houson, B. (2026). Hidden traits, visible reaction? Ability to identify criteria and reactions to an implicit test of workplace psychopathy. *Journal of Personnel Psychology*. Advance online publication. <https://doi.org/10.1027/1866-5888/a000393>
- Cook, R., Roulin, N., & Joy, K. (2024). Development, validation, and faking-resistance of an implicit measure of psychopathy in the workplace. *Human Performance*, 37(5), 245–279.
<https://doi.org/10.1080/08959285.2024.2422341>

- Davis, J., Van Bulck, L., Durieux, B. N., & Lindvall, C. (2024). The temperature feature of ChatGPT: Modifying creativity for clinical research. *JMIR Human Factors*, *11*, e53559. <https://doi.org/10.2196/53559>
- De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G. P., Ferragina, P., Tozzi, A. E., & Rizzo, C. (2023). ChatGPT and the rise of large language models: The new AI-driven infodemic threat in public health. *Frontiers in Public Health*, *11*, 1166120. <https://doi.org/10.3389/fpubh.2023.1166120>
- Dunlop, P. D., & Lievens, F. (2026). From forbid to reimagine: Employer strategies for responding to candidate GenAI use in assessment. *International Journal of Selection and Assessment*, *34*(2), e70060. <https://doi.org/10.1111/ijsa.70060>
- Edens, J. F., Buffington, J. K., Tomicic, T. L., & Riley, B. D. (2001). Effects of positive impression management on the Psychopathic Personality Inventory. *Law and Human Behavior*, *25*(3), 235–256. <https://doi.org/10.1023/A:1010793810896>
- Eisenbarth, H., Lilienfeld, S. O., & Yarkoni, T. (2015). Using a genetic algorithm to abbreviate the Psychopathic Personality Inventory–Revised (PPI-R). *Psychological Assessment*, *27*(1), 194–202. <https://doi.org/10.1037/pas0000032>
- George, A. S., George, A. S. H., & Martin, A. S. G. (2023). A review of ChatGPT AI's impact on several business sectors. *Partners Universal International Innovation Journal*, *1*(1). <https://doi.org/10.5281/zenodo.7644359>
- Google DeepMind. (2026). *Gemini 3.5 Flash* [Large language model]. <https://deepmind.google>
- Griffith, R. L., & Converse, P. D. (2012). The rules of evidence and the prevalence of applicant faking. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 34–52). Oxford University Press.

- Hare, R. D. (2003). *Manual for the Revised Psychopathy Checklist* (2nd ed.). Multi-Health Systems.
- Landay, K., Harms, P. D., & Credé, M. (2019). Shall we serve the dark lords? A meta-analytic review of psychopathy and leadership. *Journal of Applied Psychology, 104*(1), 183–196. <https://doi.org/10.1037/apl0000357>
- Landers, R. N., Sackett, P. R., & Tuzinski, K. A. (2011). Retesting after initial failure, coaching rumors, and warnings against faking in online personality measures for selection. *Journal of Applied Psychology, 96*(1), 202–210. <https://doi.org/10.1037/a0020375>
- Lee, S., Lim, S., Han, S., Oh, G., Chae, H., Chung, J., Kim, M., Kwak, B., Lee, Y., Lee, D., Yeo, J., & Yu, Y. (2024). *Do LLMs have distinct and consistent personality? TRAIT: Personality testset designed for LLMs with psychometrics* (arXiv:2406.14703). arXiv. <https://doi.org/10.48550/arXiv.2406.14703>
- Levashina, J., Weekley, J. A., Roulin, N., & Hauck, E. (2014). Using blatant extreme responding for detecting faking in high-stakes selection: Construct validity, relationship with general mental ability, and subgroup differences. *International Journal of Selection and Assessment, 22*(4), 371–383. <https://doi.org/10.1111/ijsa.12084>
- Lilienfeld, S. O., & Widows, M. R. (2005). *Psychopathic Personality Inventory-Revised*. Psychological Assessment Resources.
- Mahmut, M. K., Menictas, C., Stevenson, R. J., & Homewood, J. (2011). Validating the factor structure of the Self-Report Psychopathy Scale in a community sample. *Psychological Assessment, 23*(3), 670–678. <https://doi.org/10.1037/a0023090>

- Oostrom, J. K., De Vries, R. E., & De Wit, M. (2019). Development and validation of a HEXACO situational judgment test. *Human Performance*, 32(1), 1–29.
<https://doi.org/10.1080/08959285.2018.1539856>
- OpenAI. (2026). *GPT-5.4 Mini* [Large language model]. <https://openai.com>
- Patrick, C. J. (2010). *Operationalizing the triarchic conceptualization of psychopathy: Preliminary description of brief scales for assessment of boldness, meanness, and disinhibition*. [Unpublished manual].
- Patrick, C. J., Fowles, D. C., & Krueger, R. F. (2009). Triarchic conceptualization of psychopathy: Developmental origins of disinhibition, boldness, and meanness. *Development and Psychopathology*, 21(3), 913–938.
<https://doi.org/10.1017/S0954579409000492>
- Phillips, J., & Robie, C. (2024a). Can a computer outfake a human? *Personality and Individual Differences*, 217, Article 112434. <https://doi.org/10.1016/j.paid.2023.112434>
- Phillips, J., & Robie, C. (2024b). Hacking the perfect score on high-stakes personality assessments with generative AI. *Personality and Individual Differences*, 231, Article 112840. <https://doi.org/10.1016/j.paid.2024.112840>
- Risavy, S. D., Fisher, P. A., Robie, C., & König, C. J. (2019). Selection tool use: A focus on personality testing in Canada, the United States, and Germany. *Personnel Assessment and Decisions*, 5(1), 62–72. <https://doi.org/10.25035/pad.2019.01.004>
- Robie, C., Feeney, J. R., Rasheed, S., Bourdage, J. S., Christiansen, N. D., Dunlop, P. D., & Daljeet, K. N. (2026). Faking resistance of forced-choice and single-statement personality measures: Stress-testing with humans and generative AI. *International Journal of Selection and Assessment*, 34, e70075. <https://doi.org/10.1111/ijsa.70075>

- Robie, C., Phillips, J., Bourdage, J. S., Christiansen, N. D., Dunlop, P. D., Risavy, S. D., & Speer, A. B. (2025). Can ChatGPT outperform humans in faking a personality assessment while avoiding detection? *International Journal of Selection and Assessment*, 33(3). <https://doi.org/10.1111/ijsa.70015>
- Robie, C., Wingate, T. G., Baytalskaya, N., & Butera, H. (2026). Candidate generative AI use in pre-hire employment assessments: Self-reported incidence and the impact of warnings. *International Journal of Selection and Assessment*, 34(2), <https://doi.org/10.1111/ijsa.70056>
- Roth, L., & Klehe, U.-C. (2025). The enemy within one's own ranks: Meta-analysis on the effects of psychopathy on workplace-related behavior. *The Journal of Applied Psychology*, 110(7), 906–929. <https://doi.org/10.1037/apl0001248>
- Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Romero, P., Abdulhai, M., Faust, A., & Matarić, M. (2023). *Personality traits in large language models*. arXiv. <https://arxiv.org/abs/2307.00184>
- Smith, S. F., & Lilienfeld, S. O. (2013). Psychopathy in the workplace: The knowns and unknowns. *Aggression and Violent Behavior*, 18(2), 204–218. <https://doi.org/10.1016/j.avb.2012.11.007>
- Stevens, G. W., Deuling, J. K., & Armenakis, A. A. (2012). Successful psychopaths: Are they unethical decision-makers and why? *Journal of Business Ethics*, 105, 139–149. <https://doi.org/10.1007/s10551-011-0963-1>
- Thibault, T., & Kelloway, E. K. (2020). The dark tetrad at work. *Human Performance*, 33(5), 406–424. <https://doi.org/10.1080/08959285.2020.1802728>

- Treglown, L., & Furnham, A. (2026). AI, social desirability, and personality assessments: Impression management in large language models. *Personality and Individual Differences*, 251, 113563. <https://doi.org/10.1016/j.paid.2025.113563>
- Tu, N., Guo, F., Zhang, B., & Nie, W. (2024). *A case study on detecting ChatGPT-generated responses in personality assessments: Methods and implications*. PsyArXiv. <https://osf.io/preprints/psyar>
- Valone, A. L. Y., & Meade, A. W. (2024). Can forced-choice response format reduce faking of socially aversive personality traits? *Journal of Personality Assessment*, 106(6), 819–831. <https://doi.org/10.1080/00223891.2024.2326893>
- Walker, S. A., Double, K. S., Birney, D. P., & MacCann, C. (2022). How much can people fake on the dark triad? A meta-analysis and systematic review of instructed faking. *Personality and Individual Differences*, 193, Article 111622. <https://doi.org/10.1016/j.paid.2022.111622>
- Zhou, J., Lu, T., Mishra, S., Brahma, S., Basu, S., Luan, Y., Zhou, D., & Hou, L. (2023). *Instruction-following evaluation for large language models*. arXiv. <https://arxiv.org/abs/2311.07911>