

The performance of large language models on quantitative and verbal ability tests: Initial evidence and implications for unproctored high-stakes testing

Louis Hickman¹ | Patrick D. Dunlop²  | Jasper Leo Wolf³

¹Department of Psychology, Virginia Tech, The Wharton School of the University of Pennsylvania, Blacksburg, Virginia, USA

²Future of Work Institute, Faculty of Business and Law, Curtin University, Perth, Western Australia, Australia

³Arctic Shores, London, UK

Correspondence

Louis Hickman, Department of Psychology, Virginia Tech; The Wharton School of the University of Pennsylvania, Blacksburg, VA, USA.

Email: louishickman@vt.edu

Abstract

Unproctored assessments are widely used in pre-employment assessment. However, widely accessible large language models (LLMs) pose challenges for unproctored personnel assessments, given that applicants may use them to artificially inflate their scores beyond their true abilities. This may be particularly concerning in cognitive ability tests, which are widely used and traditionally considered to be less fakeable by humans than personality tests. Thus, this study compares the performance of LLMs on two common types of cognitive tests: quantitative ability (number series completion) and verbal ability (use a passage of text to determine whether a statement is true). The tests investigated are used in real-world, high-stakes selection. We also examine the performance of the LLMs across different test formats (i.e., open-ended vs. multiple choice). Further, we contrast the performance of two LLMs (Generative Pretrained Transformers, GPT-3.5 and GPT-4) across multiple prompt approaches and “temperature” settings (i.e., a parameter that determines the amount of randomness in the model's output). We found that the LLMs performed well on the verbal ability test but extremely poorly on the quantitative ability test, even when accounting for the test format. GPT-4 outperformed GPT-3.5 across both types of tests. Notably, although prompt approaches and temperature settings did affect LLM test performance, those effects were mostly minor relative to differences across tests and language models. We provide recommendations for securing pre-employment testing against LLM influences. Additionally, we call for rigorous research investigating the prevalence of LLM usage in pre-employment testing as well as on how LLM usage affects selection test validity.

KEYWORDS

artificial intelligence, chatbots, cognitive ability testing, generative pretrained transformer, large language models

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Authors. *International Journal of Selection and Assessment* published by John Wiley & Sons Ltd.

Practitioner points

- Job candidates may use large language models like ChatGPT to complete ability tests on their behalf, but we currently know little about how well these models perform on commercial cognitive ability tests.
- OpenAI's (free) Generative Pretrained Transformers (GPT)-3.5 and (paid subscription) GPT-4 models both performed very poorly on a quantitative ability test.
- GPT-4 achieved results above the 90th percentile on the verbal ability test, and GPT-3.5 scored at approximately the 60th percentile.
- Temperature settings did not substantially affect the performance of the large language models and different prompt approaches tended not to, with few exceptions.

Large language models (LLMs), such as Claude, Bard, LLaMA, and Generative Pretrained Transformers (GPT) that power ChatGPT, recently emerged as powerful artificial intelligence (AI) tools that are highly accessible and offer a huge range of potential use cases. Of particular concern to personnel assessment and selection is the case where a job candidate uses an LLM to assist with, or complete on their behalf, an unproctored assessment. Our study aims to advance our understanding of how well LLMs perform on unproctored cognitive ability tests used in high-stakes selection, thus providing important insight into the potential threat of LLMs to unproctored pre-employment assessments.

Upon its release to the general public, OpenAI detailed GPT-4's performance on a range of program entry and professional licensure assessments, including the graduate record examination (GRE), SAT, law SAT (LSAT), and the Uniform Bar Exam (OpenAI, 2023). In many cases, the LLM responses scored at or above the 90th percentile, compared to human test takers. Independent research teams have subsequently examined the performance of LLMs in relation to, for example, medical situational judgment tests (Borchert et al., 2023), medical knowledge and "soft skills" assessments (Brin et al., 2023), and both single-stimulus and forced-choice personality assessments (Phillips & Robie, 2024). These studies' findings align with OpenAI's report: advanced LLMs outscore most human test takers on several types of tests. This raises new concerns about unproctored employment testing, given that tests that are more difficult to fake—like cognitive ability tests—are now potentially fakeable by anyone with access to an LLM.

To investigate the extent to which job applicants may be able to cheat using LLMs, we investigate GPT's performance on two commercially available cognitive ability tests used for personnel selection. Given the rapid growth in the use of LLMs (Porter, 2023) and that, at the time of writing, LLM usage was highest among people aged 34 and under (Similar Web, 2023), the tests we examined are used primarily as remote, unproctored tests for screening recent college graduates (Similar Web, 2023). Backed by research suggesting that unproctored pre-employment testing does not substantially inflate scores or decrease criterion-related validity (e.g., Beaty et al., 2011; Lievens & Burke, 2011), many employers now routinely

assess candidates remotely. However, if LLMs score well on validated cognitive ability tests, our field may need to reassess concerns about unproctored testing in personnel selection.

Our investigation provides three major contributions. First, we compare LLM performance on two common types of cognitive tests: quantitative ability and verbal ability. Further, we examine performance across different response formats (i.e., open-ended vs. multiple choice). As we explain in detail below, the foundations of LLMs imply that certain test characteristics (e.g., subject matter, response format) will likely affect their vulnerability to LLM-based cheating.

Second, we contrast the performance of two LLMs, namely GPT-3.5 and GPT-4. Such a comparison is critical to understanding the capacity of LLMs in principle, and the performance of LLMs in practice as candidates would use them. Importantly, at the time of writing, GPT-3.5 is freely available via ChatGPT and may therefore be most representative of the typical use case by candidates. Its model architecture and training data, however, is considerably smaller than that of GPT-4, which requires a paid subscription to access via ChatGPT (at time of writing, US \$20/month), or per-token access via the application programming interface (API).¹ The subscription may prove a barrier to entry for candidates with fewer financial resources or those simply preferring not to pay the fee. Further, the API is likely a major barrier to access for candidates without requisite technical skills.

Third, we also examine the impact of different prompt approaches. As we explain in detail below, it is now well established that the nature, depth, and quality of responses that a LLM generates is dependent on the content of the input provided by the user: the "prompt" (e.g., chain-of-thought; Wang et al., 2022; Wei et al., 2022). It remains less clear, however, precisely if and how prompt approaches affect LLM scores on cognitive tests.

1 | LARGE LANGUAGE MODELS

Large language models are one of the most recent, significant developments in the field of natural language processing (NLP). LLMs operate similarly to the predictive text function on a cell phone, email

client, or word processor but involve larger model architectures and more training data. These LLMs utilize deep learning—or neural networks with many hidden layers and nodes per layer—and are trained on huge natural language text corpora (i.e., collections of documents, websites, and books) to conduct next word prediction (e.g., Brown et al., 2020). The training process involves giving the LLM text, having it predict the next word, then updating the model's weights to improve the accuracy of the prediction, and repeating that process an extremely large number of times. Some recent models have additionally been trained with reinforcement learning using human and AI feedback on previous models' outputs (Ray, 2023). This involves updating the model's weights to make its predictions more like those that users rated positively, while making predictions less like the predictions that users rated negatively.

LLMs gain a surprisingly robust probabilistic representation of the grammar, syntax, and semantics of natural language. When LLMs generate new text (e.g., as occurs when querying ChatGPT), responses are generated one word at a time: the first word is generated, then it is appended to the model's input and the second word is generated, which is then appended to the model input, and so on, until the response is complete. Given the nature of LLMs, their representation of semantics and their predictions (i.e., responses) are strongly influenced by context, which is determined by the training data and the user-provided prompt.

The first LLM from OpenAI was GPT-1, which used a transformer architecture. GPT-1's neural network included 117 million parameters (i.e., weights connecting neurons and a constant "bias" added at each neuron; Radford et al., 2018). GPT-2 used a much larger neural network, with 1.5 billion parameters, and it was trained on more data than GPT-1 (Radford et al., 2019). As a result, GPT-2 generated text was much more human-like than text from GPT-1. This evolution continued with GPT-3, which included 175 billion parameters in the neural network and, again, used more training data—300 billion tokens (a token is the basic unit that LLMs process, a word or subword—on average, one word is approximately 0.75 tokens; Brown et al., 2020). GPT-3 also had a larger context window than earlier models at 2049 tokens (3.5 had a context window of 3096 tokens, which is the maximum length of content the model can accept when generating responses). GPT-4 grew further, with researchers estimating that the model has 1.76 trillion parameters in its neural network and that it was trained on 14 trillion tokens (Schreiner, 2023). The minimum context window of the GPT-4 class of models is 8192 tokens. GPT-4 was additionally trained on human-generated feedback on output sequences generated through ChatGPT, and it can accept images or text as input (Ray, 2023). As of April 2024, the LMSYS Chatbot Arena Leaderboard (which asks users to make "blind" comparisons of the quality of two LLMs' outputs; <https://chat.lmsys.org/?leaderboard>) suggests that several other LLMs perform comparably to GPT-4, including Anthropic's CLAUDE 3, Mistral AI's Mistral, and Google's Bard/Gemini Pro.

Despite the impressive performance of GPT-4 and other LLMs on many tasks, they are probabilistic in nature and lack consciousness. They "learn" the correlations between different tokens very

well, leading them to display behaviors that some people interpret as human-like understanding (Mitchell & Krakauer, 2023). However, the brittleness of these systems illustrates that they do not have any understanding of what is being asked of them nor what they reply. Instead, they are merely neural networks that use statistics to predict the next, most likely word, given the inputs.

It is well known that LLM outputs are imperfect and can include misleading or outright false statements. Some have termed errors by an LLM as "hallucinations" (Maynez et al., 2020) or "confabulations" (Ali et al., 2023). When trained on language, the models get access to form, but they do not get access to semantics (Bender et al., 2021). Thus, these models cannot lead to a more general AI that truly understands its inputs, outputs, and the consequences of its actions (Floridi & Chiriatti, 2020). This suggests that additional increases in neural network and training data size will provide asymptotic benefits unless the nature of the models is fundamentally altered. In other words, continuing to simply increase the number of parameters, the size of the context and training data will likely lead to an asymptote in performance gains, and it may indeed be that the current, best-performing models are already close to that point. Altogether, while LLMs have many powerful use cases and can be beneficial for completing a variety of tasks (e.g., Budhwar et al., 2023), their performance on any given task is a function of their architecture (i.e., the size and shape of the neural network; the optimization function used), the quality, diversity, and size of their training data, and the prompt inputted to them.

2 | UNPROCTORED COGNITIVE ABILITY TESTING

Seminal, influential meta-analytic work in industrial and organizational psychology by Schmidt and Hunter (1998) concluded that cognitive ability tests are among the strongest predictors of job performance across a variety of occupations (while updated meta-analyses suggest smaller validities, they still rank among the strongest predictors; Sackett et al., 2022). Historically, cognitive ability testing would be a proctored, face-to-face "paper-and-pencil" activity. However, over the past three decades, technology enabled the development of unproctored internet testing, affording flexibility to candidates and scale to employers.

Initially, the pivot to remote, unproctored testing practices raised concerns among academics, particularly around the potential for candidates to "cheat" in an unproctored environment by receiving assistance or having an acquaintance take the test on their behalf (e.g., Tippins et al., 2006). Indeed, cognitive ability test scores are, on average, slightly higher when unproctored than when proctored (Steger et al., 2020). Other research, however, suggests that unproctored tests function similarly to proctored versions (Templer & Lange, 2008), and that base rates of cheating and suspicious scoring patterns on unproctored tests are low (Arthur et al., 2010; Kantrowitz & Dainis, 2014; Nye et al., 2008). In other words, even if some candidates were receiving assistance completing unproctored

tests or entirely outsourcing test completion to others, detrimental effects on the integrity of the selection system across the candidate pool are considered negligible. Altogether, to this point, cognitive test performance would seem to represent a (largely) honest, difficult-to-fake, signal of the candidates' ability (Bangerter et al., 2012).

On the one hand, the apparent absence of cheating effects on test integrity might signal a general reluctance of candidates to cheat, implying that even the availability of LLMs should not give cause for further concern about unproctored internet testing. On the other hand, research has shown that a nontrivial proportion of candidates are willing to exaggerate or fabricate information when completing noncognitive assessments (e.g., resumes, biodata questionnaires, interviews, personality inventories) (Donovan et al., 2003; Hu & Connelly, 2021; Levashina & Campion, 2007; Levashina et al., 2009), implying that a meaningful proportion of candidates may be willing to cheat on ability tests when barriers to doing so are removed. Indeed, in the context of personnel selection, signaling theory suggests that candidates have a clear motivation to attempt to send inaccurate but positive signals to employers (Bangerter et al., 2012).

Importantly, before LLMs, a candidate wishing to cheat on an ability test still must find an able, willing, and available volunteer. For example, if a candidate of moderate ability outsources their test taking to, or receives coaching from, somebody also of moderate ability, this candidate may not benefit from that cheating attempt. This differs from self-report personality tests, where cheating (or faking) is relatively simple to achieve, because self-report personality tests ask people to *report* their standing on a trait, whereas cognitive ability tests require people to *demonstrate* their standing on the construct (Chamorro-Premuzic & Furnham, 2005). In terms of signaling theory, priors to LLMs, the costs of cheating on a cognitive ability test (i.e., time taken to identify and convince an intelligent acquaintance to be a surrogate test-taker) are much higher than those for a self-report assessment (Bangerter et al., 2012).

By contrast, if LLM outputs reliably achieve high scores on cognitive ability tests, then a significant practical barrier to cheating is effectively removed, and candidates who cheat may displace candidates who do not. In other words, the costs of cheating are drastically reduced (Bangerter et al., 2012). Thus, the public availability and accessibility of LLMs has altered the landscape of unproctored testing because anyone could potentially use an LLM to cheat on unproctored cognitive ability tests.

3 | THE TEST-TAKING CAPABILITIES OF LARGE LANGUAGE MODELS

There remains considerable uncertainty with respect to the problem-solving capabilities of LLMs. Published research suggests that LLMs often achieve high scores on tests that comprise extensive verbal information. As noted above, advanced LLMs appear to achieve higher scores than the majority of humans on many knowledge-based assessments (OpenAI, 2023), certain situational judgement tests (Arctic Shores, 2023b; Borchert et al., 2023), and personality

questionnaires (if prompted to; Arctic Shores, 2023a; Phillips & Robie, 2024). Further, Elyoseph et al. (2023) found that ChatGPT (GPT-3.5) outperformed most humans on an emotional awareness assessment comprising items with text descriptions of situations that required participants to identify an emotional state. By contrast, Groza (2023) presented ChatGPT (GPT-3.5) with 100 logical and numerical reasoning problems of various types and found it correctly answered only 16 of them and demonstrated many logical fallacies in its reasoning. Similarly, Mitchell, and Palmarini, and Moskvichev (2023) found that GPT-4 and GPT-4 Vision (an OpenAI model that can receive visual inputs in addition to text) performed poorly relative to humans on a set of abstract reasoning problems.

Digging deeper into OpenAI's (2023) reported results reveals that the version of the test completed was often an old, defunct version of the test or a practice test. Given that OpenAI no longer reveals the full details of their training data and process, it is impossible to know whether questions from these tests were in the LLM's training data—a concern since LLMs sometimes reproduce their training data word-for-word (e.g., Nasr et al., 2023)—which is one potential cause of the large variance in GPT's performance on different tests. Further, LLM developers also rarely provide details about the specific prompts used, which raises the risk of *p*-hacking or selective reporting if, for example, they only disclose results for the single prompt that generated the best results, despite trying many alternatives.

Although early research identified several areas where LLMs often perform better than many humans, there remain gaps in our understanding. First, very few tests examined in prior research are actually used in workplace personnel selection (we note that OpenAI reported performance on several tests used in academic selection, albeit using old versions of the exams or practice exams; see OpenAI, 2023), aside from specialized medical tests. This distinction is vital, because tests used in high-stakes selection settings may differ in several ways from research-oriented tests or tests used for academic selection. In our investigation, we examine two types of tests used for high-stakes selection, quantitative ability and verbal ability, and we compare LLM performance across the two types of tests.

Research Question 1. *How do LLMs perform on quantitative and verbal ability tests used in high-stakes settings?*

Second, the effectiveness of LLM outputs can be highly dependent on the prompt that the model is given, as the prompt influences the probabilities of certain words appearing. An emerging line of research is focused on identifying prompt approaches that increase the quality of an LLM's outputs (often termed "prompt engineering;" see Chen et al., 2023, for a review). This research is still nascent, and as such, inconsistencies in nomenclature and definitions are rife. Nonetheless, several formal prompt approaches have been proposed, along with circumstances under which certain styles are expected to be more effective. Importantly, we note that prior research into the effectiveness of LLMs in completing assessments has primarily focused on finding *one* prompt that seemed to work well, then using that prompt for generating all responses.

There are reasons to suspect, however, that candidates using LLMs to assist with a test may employ a variety of different prompt approaches. First, the popular literature on LLMs and Generative AI (e.g., Metcalfe, 2024; Morton, 2024), as well as LLM websites (e.g., Anthropic, 2024) provides extensive discussion and description of prompt engineering. Second, many online sources offering specialist career advisory services (e.g., resume writing), and more general online news sources (e.g., Forbes, Medium), and social media sites (e.g., YouTube, Reddit, TikTok) provide advice to job seekers seeking to exploit LLMs to assist with their application materials, and in doing so, highlight the importance of effective prompting. For example, a search on Google conducted by one of the authors for “using ChatGPT to improve your resume” yielded many hits comprised almost entirely of links to articles or videos from the types of sources just mentioned, and most containing specific references to prompting. And third, at the time of writing, myriad short courses on “prompt engineering” were available on online educational services such as LinkedIn Learning (23 courses), Udemy (680 courses and videos), and Coursera (273 courses). While not every LLM-using candidate will be aware of prompting approaches, altogether, it seems likely that many candidates will have at least formed a hypothesis that, or about how, certain prompts will affect the LLM's performance. It is impossible to examine all possible prompt approaches, however, it is important to investigate whether several of the current, more popular and intuitive approaches might influence a LLM's performance on tests (Mitchell & Palmarini, & Moskvichev, 2023).

First, we considered two very basic prompting approaches that provide minimal context and no special instruction. The first involves copy-pasting the test instructions and question as input to the LLM. In our investigation, this was termed “Base Prompting.” Similarly, we explored an even more minimalistic approach that only included the portion of the instructions that asked the focal question (as detailed in the Supporting Information S1: Method and the Online Supplement), which we referred to as “Bare Bones.” Additionally, we investigated three relatively well-established but more complex prompt approaches that candidates who have investigated prompt engineering might encounter. These were “chain-of-thought,” “persona/role-based,” and “vocalize and reflect.” Chain-of-thought involves a prompt that includes an example of a similar question, the correct answer, and step-by-step rationale (i.e., chain-of-thought) for how to solve the problem (Wei et al., 2022). Persona/role-based prompt approaches involve giving the LLM a “persona” so that it will respond more like a certain type of person (Xu et al., 2023). Generally, the persona is a subject matter expert on the topic, with the logic being that by asking the LLM to adopt the mindset of the expert, it will respond more similar to such experts. For example, when completing the verbal ability test, we instructed it, “You are an experienced corporate analyst at a large organization, known for your meticulous attention to detail and your ability to make accurate judgments based on documents and reports.” Vocalize and reflect involves asking the LLM to develop an initial response, critique it, and then develop a revised response based on its own critique—a form of self-learning (Shinn et al., 2023). The hope is that by asking the LLM to critique its own initial response, it will arrive at an improved response. Finally, we also investigated a fifth strategy which we termed “method

explained.” Method explained instructs the LLM to explain the methodology it will use for correctly completing the question before answering (Kojima et al., 2023). In real applicant settings, this could give applicants the opportunity to advise the LLM on how to think through the problem more effectively.²

Research Question 2. *Do different prompt approaches influence LLM performance on cognitive ability tests used in high-stakes settings?*

Current evidence suggests that GPT-4 outperforms GPT-3.5 in a variety of tasks (OpenAI, 2023; Phillips & Robie, 2024). One reason that this is concerning is that applicants with fewer financial resources may not be able to afford to access GPT-4 and may, instead, rely on the free ChatGPT which uses GPT-3.5. Thus, while examining GPT-3.5's performance may not provide insight into a LLM's full potential, it provides insight into how the majority of candidates relying on a LLM might perform on tests. Indeed, as discussed above, GPT-4 has two advantages over the earlier version. First, it is trained on more data. Second, it is a larger neural network. These differences, coupled with clear evidence that GPT-4 outperforms GPT-3.5 in a variety of ways, brings us to our first hypothesis.

Hypothesis 1. GPT-4 will perform better than GPT-3.5 on cognitive ability tests used in high-stakes settings.

Finally, we note that, to our knowledge, prior research has rarely considered the temperature setting in LLMs.³ The temperature setting affects how much randomness is added to the model when it is generating next word predictions. Higher temperatures introduce more randomness, while lower temperatures provide more consistent responses. At time of writing, ChatGPT defaults to a temperature value of 0.7 for both GPT-3.5 and GPT-4, whereas this can be varied when interacting with GPT through the API. Even at a temperature of 0, however, there is some randomness due to the nondeterministic nature of the LLM models. Given that higher temperatures are thought to be useful for creative tasks, it is unclear how temperatures might affect LLM performance on cognitive ability tests. Although, in principle, we could examine the LLM's performance at any temperature setting, in line with the comparison above, we aimed to compare the typical candidate use case (i.e., the current default temperature = 0.7) to a use case where we would expect the LLM to perform most consistently (temperature = 0).

Research Question 3. *Does the temperature setting influence LLM performance on cognitive ability tests used in high-stakes settings?*

4 | METHOD

Representative Python code for generating the results and the combined results file are available on OSF: https://osf.io/pja32/?view_only=af8e0fdf7aef40eab3192f65b35618d5.

4.1 | Procedure

We used several prompt approaches to generate responses from GPT-3.5 and GPT-4 to quantitative ability and verbal ability tests used in high-stakes selection settings. In interacting with the LLM, we used the OpenAI API. Using the API was necessary given the scale of our investigation: we collected over 100 sets of test responses from LLMs for the quantitative ability test and over 40 sets of responses for the verbal ability test. Specifically, we generated responses in $2 \times 2 \times 6$ (24) experimental conditions for each test: two versions of GPT (3.5 vs. 4), two temperatures (0.0 vs. 0.7), and six prompt approaches (base, bare bones, chain-of-thought, persona, vocalize and reflect, and method explained), and we generated two sets of responses in each condition for the quantitative ability test, a multiple-choice version of the quantitative ability test (described below), and the verbal ability test. Additionally, we generated one set of responses in each condition for the modified version of the multiple-choice quantitative ability test, and for the original tests, we generated a third set of responses when the first two trials disagreed by four or more points. We then took the average of all trials within each condition and rounded to the nearest integer before calculating percentile scores.

In all cases, each test question and prompt were presented to the LLM independently of the other questions. In other words, each test question was treated as a new conversation, which prevents the context created by prior test questions from influencing the output for subsequent questions. We generated all responses between October 31, 2023 and December 1, 2023. We compared the scores obtained from the GPT LLMs to scores obtained by humans in high-stakes settings. Below, we provide details about the two tests, the high-stakes human responses to which we compared the LLM-based scores, and the LLMs and prompts.

4.2 | Quantitative ability test

The quantitative ability test is a published commercial test used in personnel selection. It consisted of 20 open-ended number series items. In these items, rather than every question asking for the final item in the sequence, the location of the missing item varied across questions. The test instructs test takers to determine the missing number in a numerical sequence and gives an example. That example item is: (2, 4, 6, None, 10, 12), with the correct answer being "8." In the test items, some questions had two sequences, and the test taker must determine both the function for the sequence and which sequence applies to the missing number. The instructions clarify this, "There may be one or multiple rules governing the pattern in each sequence. The answer will always be an integer, and may also be negative (-)." For example, possible items (but not items actually on the test) would include, "(2, 5, 11, None, 35, 71) (answer: 23)," "(3, 4, 8, 10, 13, None)" (answer: 16), and "(2, 12, 17, 68, 71, None)" (answer: 142). In high-stakes settings, the test's time limit is 17 min.

The test manual reports the test's internal consistency $\alpha = 0.70$ and test-retest reliability = 0.62. The test converged $r = 0.58$ and $r = 0.62$ with other commercial numerical reasoning tests. Additionally, the test converged $r = 0.61$ with a number-letter series test and $r = 0.60$ with a 3D rotation test from the International Cognitive Ability Resource (The International Cognitive Ability Resource Team ICAR, 2014).

We found that the LLMs struggled to provide accurate responses to the open-ended format of this test. Thus, we also examined the LLM performance on a multiple-choice version of the test. To develop the multiple choices, we included the correct response, a response from a pattern that was not for that missing item, a response from an incorrect (but plausible) pattern, and "none of these."

Additionally, given that the LLMs struggled with the multiple-choice version of the test, we also examined the LLM performance on a modified version of the multiple-choice version of the test. Specifically, we edited all items so that the missing number came last in the numerical sequence. For example, revising the example item above in this way would make the sequence, (2, 4, 6, 8, 10, None). We considered this important for understanding whether the LLMs perform better when all relevant information can be gleaned from moving left-to-right, as in English language text. No psychometric properties are available that describe the multiple-choice versions of the test, given that these versions were created solely to see if that would improve the LLM test scores.

4.2.1 | Human norm scores for quantitative ability

The norm scores were derived from a sample of 9253 real-world job applicants. All applicants applied to jobs targeted toward soon-to-graduate with Bachelor's degree or recent graduates in the United Kingdom. The jobs they applied to were almost all in information technology and other professional services (99.41%). Applicant age and ethnicity were unavailable for this sample. 49.04% of applicants reported their gender as male, 32.31% did not report their gender, 18.57% reported being female, and 0.15% reported other.

4.3 | Verbal ability test

The verbal ability test, also a commercial test used for selection, consisted of 24 multiple-choice verbal reasoning items. Each item displayed a passage of text and then asked test takers to determine whether a statement about the passage is true, false, or cannot say. The key passages from the example item in the test instructions are, "All current employees are either team members, team leaders, or department managers. ... All team members have a designated mentor, but they are not permitted to act as a mentor for another employee, and all department managers have one designated mentee." The statement to evaluate is, "All employees are a mentor, a mentee, or both." The answer is "Cannot Say," because team

leaders are not mentioned in the mentor/mentee assignments. In high-stakes settings, the test's time limit is 18 min.

The test manual reports the test's internal consistency $\alpha = 0.72$ and test-retest reliability $= 0.59$. The test converged $r = 0.65$ and $r = 0.74$ with other commercial verbal reasoning tests. Additionally, the test converged $r = 0.58$ with a verbal reasoning test from the International Cognitive Ability Resource (The International Cognitive Ability Resource Team ICAR, 2014).

4.3.1 | Human norm scores for verbal ability

The norm scores were derived from a sample of 38,896 real-world job applicants. The vast majority of applicants applied to jobs targeted toward applicants who were soon-to-graduate or had recently graduated with Bachelor's degrees (98.67%), and applicants' mean age was 22.67 (SD = 4.25). Most of the data were from applicants to jobs in the United Kingdom (60.79%), with the remaining applying to jobs in the United Arab Emirates (25.73%), Australia (12.70%), and other locations (0.78%). The jobs they applied for were in banking, finance, human resources, and other professional services (32.24%), science, technology, engineering, and math (STEM; 29.11%), the public sector (7.87%), or other (30.79%). Most applicants did not voluntarily report their gender (45.37%), while 34.37% reported being male, 20.17% reported being female, and 0.09% reported other. The vast majority of applicants did not voluntarily report their race/ethnicity (86.27%), while those who did were white (8.29%), Asian (3.47%), Black/African/Caribbean (0.72%), multiple (0.71%), or other (0.55%).

4.4 | Large language models

ChatGPT has popularized the family of GPT large language models. The "generative" refers to its capacity for creating new content. "Pretrained" refers to the extensive amount of data it has been trained on to do next word prediction. "Transformer" refers to aspects of the neural network's architecture (Radford et al., 2018). We compared the performance of GPT-3.5, as implemented in the OpenAI API as gpt-3.5-turbo, and GPT-4, as implemented in the OpenAI API as gpt-4. GPT-3.5 is freely available via ChatGPT, while GPT-4 can be accessed via the API or the paid version of ChatGPT.

4.5 | Prompt approaches

We utilized six prompt approaches: base, bare bones, chain-of-thought, persona, vocalize and reflect, and method explained. The prompt approaches used are listed in Supporting Information S1: Table S1 for the quantitative ability test and Supporting Information S1: Table S2 for the verbal ability test. The *base* prompt approach involved providing the LLM with (a) the instructions from the test, (b) the item along with the response options, and (c) instructions to work

out the answer step-by-step and provide its final answer.⁴ The *bare bones* prompt approach merely pasted in the response instructions from each test (i.e., for the quantitative ability test, "What is the missing number in the sequence below," and for the verbal ability test, "Is the following statement true, false, or cannot say?"). This approach was intended to be similar to what naïve test takers might do. *Chain-of-thought* involved providing the LLM with (a) the instructions from the test, (b) an example item, (c) a step-by-step process for reasoning through and correctly answering the example item as well as the correct answer, and (d) elements (b) and (c) from the base prompt. *Persona* involved specifying a persona to inhabit before the test instructions. For the quantitative ability test, we used the persona of a math teacher who excels at pattern recognition and explaining math solutions in a logical, easy-to-understand way, and for the verbal ability test, we used the persona of an experienced corporate analyst known for their attention to detail and accurate judgments from documents. For *vocalize and reflect*, we modified the base prompt by adding content after the test instructions that instructed the LLM to provide its initial judgment and explain its reasoning, then review its initial response to identify possible improvements, before providing a final answer with an explanation for its reasoning. Finally, *method explained* expanded on the base prompt approach by instructing the LLM to outline in detail the methodology for accurately completing such tests.⁵

5 | RESULTS

Research Question 1 focused on the LLMs' performance on the quantitative ability test versus the verbal ability test. The average results across runs and other conditions are reported in Table 1 and illustrated in Figure 1 (for the verbal ability test and only the open-ended [original] version of the quantitative test). The results broken down by conditions are reported in Supporting Information S1: Tables S3 (quantitative) and S4 (verbal).

The LLMs performed substantially better on the verbal ability test (average 69.97 percentile) than on the quantitative ability test (average 8.38 percentile). This finding held even when the quantitative ability test was in a multiple choice format (like the verbal ability test; those results are reported in Supporting Information S1: Table S5; average 15.18 percentile, compared to humans responding to the open-ended (original) version of the test), and when we shifted the missing item to be the final item in the numerical sequence (Supporting Information S1: Table S5 under 'missing number last'). Although the LLMs scored considerably higher on the quantitative ability test when all the missing items were the final items in the numerical sequences (average 28.81 percentile, again compared to humans responding to the open-ended [original] version of the test), they still performed worse, on average, compared to their performance on the verbal ability test.

Research Question 2 regards the performance of different prompt approaches. On average across both tests, the bare bones prompt approach provided the worst scores overall. On the

TABLE 1 Average performance of Large language models on quantitative and verbal ability tests.

GPT-3.5					GPT-4				
	Quantitative		Verbal			Quantitative		Verbal	
	Score	Percentile	Score	Percentile		Score	Percentile	Score	Percentile
Overall	3.46	4.69	12.39	45.39	Overall	6.25	12.06	19.46	94.55
Prompt					Prompt				
Base	4.50	9.29	13.17	54.24	Base	4.25	6.70	19.00	94.55
Bare Bones	3.50	6.70	7.00	8.96	Bare Bones	5.25	9.29	17.50	90.58
COT	3.75	6.70	14.00	62.90	COT	7.50	18.73	22.00	99.57
Persona	3.75	6.70	13.50	62.90	Persona	6.75	15.14	20.00	97.23
V&R	2.75	4.69	12.92	54.24	V&R	8.00	18.73	18.75	94.55
ME	2.50	4.69	13.75	62.90	ME	5.75	12.06	19.50	97.23
Temperature					Temperature				
0.0	2.92	4.69	12.42	45.39	0.0	5.75	12.06	19.17	94.55
0.7	4.00	6.70	12.36	45.39	0.7	6.75	15.14	19.75	97.23

Note: Score is the mean score achieved across multiple trials. All quantitative test results are for the open-ended (original) version of the test. Overall = average score achieved for specified GPT model across all prompts and temperatures. Initially, two trials were used for each LLM-prompt-temperature triplet, and a third was added when the first two trials disagreed by four or more points. Prompt scores are averaged across the trials and two temperatures. Temperature scores are averaged across the trials and six prompts. Percentiles calculated after rounding mean score to the nearest integer based on responses to the original tests.

Abbreviations: COT, Chain of thought; ME, Method Explained; V&R, Vocalize and Reflect.

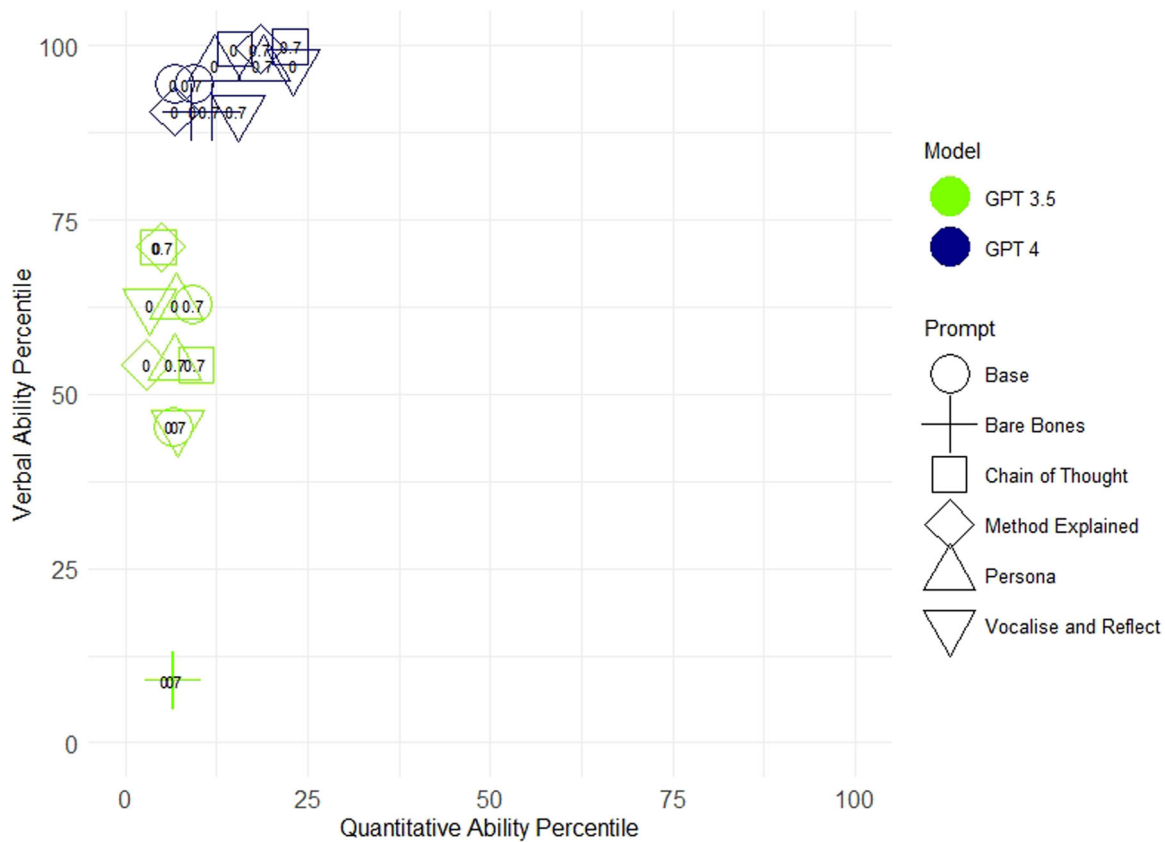


FIGURE 1 Percentiles Achieved by GPT Across Prompts, Models, and Temperatures. Temperatures indicated with text on the plot points. All quantitative test results are for the open-ended (original) version of the test. Percentiles calculated after rounding mean score to the nearest integer based on responses to the original tests.

quantitative test, bare bones performed comparably to the other prompt approaches, but on the verbal ability test—particularly with GPT-3.5—bare bones performed much worse than the other prompt approaches (average score of 7 compared to average score 12.92 or above for the other prompt approaches). Bare bones also performed worse than the other prompt approaches with GPT-4 on the verbal ability test, but the differences were much smaller. On average across both tests, the chain-of-thought prompt approach yielded the best scores. The average differences between different prompt approaches were not large (except with the bare bones prompt approach)—for the quantitative ability test, the range of average values for GPT-3.5 was 2.00 and for GPT-4 was 3.75. For the verbal ability test, the range of average values for GPT-3.5 was 7.00 (driven mainly by the low score from the bare bones prompt approach) and for GPT-4 was 4.50. Thus, with the exception of the bare bones prompt approach, the scores and percentiles achieved by different prompt approaches were highly similar.

Supporting Hypothesis 1, GPT-4 outperformed GPT-3.5 in both the quantitative and verbal ability test. On average across prompt approaches and temperatures, GPT-4 scored almost twice as high as GPT-3.5 on the quantitative ability test, although in both cases, the scores fell below the 20th percentile (on average across prompt approaches and temperatures). On average, GPT-4 scored 6.38 points (almost 50%) higher on the verbal ability test than GPT-3.5, which resulted in averaging in the 95th percentile for GPT-4 as compared to the 45th percentile for GPT-3.5.

Research Question 3 regards the influence of the temperature setting on the LLMs' scores. Overall, on average, a temperature of 0.7 provided slightly higher scores (mean = 10.72) on both tests than a temperature of 0.0 (mean = 10.06). The difference was just over 1 point for the quantitative ability test (mean difference = 1.04). And the difference was small for the verbal ability test (mean difference = 0.26). Thus, a nonzero temperature provided small improvements in test performance.

We also examined which items the LLMs tended to get correct. To do so, we used the human norm score data to calculate the percentage of correct responses for each item. We used this as a measure of item difficulty (i.e., item difficulty as the inverse of the proportion of test takers who answered the item correctly). For verbal ability, item difficulty correlated $r = -0.24$ and -0.35 , respectively, with the average GPT-3.5 and GPT-4 item scores (across temperature and prompt approaches). For quantitative ability, item difficulty correlated $r = -0.40$ and -0.50 , respectively, with the average GPT-3.5 and GPT-4 item scores. In particular, for the quantitative ability test, we found that GPT performed poorly on items that involved alternating sequences (i.e., the sequence alternates between addition/subtraction and multiplication/division), especially when the mathematical operation increased or decreased in value throughout the sequence. Further, as demonstrated above, it performed worse on items where the missing number was not the final number in the sequence. GPT performed well on items with relatively simple, single sequences (e.g., multiplying each number in the sequence by a constant value). Overall, the LLMs were more

likely to provide correct responses to easier items and less likely to do so for more difficult items.

6 | DISCUSSION

In an era of widespread unproctored testing and LLM availability, it is critical to understand the extent to which candidate assessment may be disrupted by this new technology. Accordingly, in the first known study of its kind, we examined the performance of two LLMs on commercial quantitative and verbal ability tests used for personnel selection. To these ends, we evaluated the performance of two popular LLMs (the freely available GPT-3.5 and the subscription based GPT-4) with six different prompt approaches and two temperature settings. We found, first, that LLMs performed substantially better on the verbal ability test than on the quantitative ability test. Second, we found that the GPT-4 model vastly outperformed GPT-3.5 on the verbal assessment but performed only slightly better on the quantitative test. Third, we found the different prompt approaches we tested were not associated with large differences in performance, except for the bare bones approach on the verbal ability test with GPT-3.5, which produced notably worse results. And fourth, we found slight improvements in performance when increasing the models' temperature from the minimum of zero to their default value of 0.7.

6.1 | Implications for understanding LLMs

The impressive performance of GPT-4 LLM on the verbal test is consistent with findings from other research that has demonstrated GPT-4's capability in completing a variety of assessments involving a large proportion of verbal content (Brin et al., 2023; OpenAI, 2023). The strong performance of this model has been observed on assessments involving veridical truths (i.e., knowledge and aptitude; Brin et al., 2023) and subjective or socially constructed "truths" (e.g., personality, emotional intelligence, situational judgment; Borchert et al., 2023; Phillips & Robie, 2024). By contrast, the GPT-3.5 model, which is trained on a smaller data set and has fewer model parameters than GPT-4, performed only slightly better than the median member of the human norm group. Based on these results, only GPT-4 (and perhaps similarly performing models) hold potential to threaten the integrity of unproctored verbal ability assessments, whereas applicants relying on the free GPT-3.5 model are unlikely to achieve stand-out results. One potential implication for this pattern of results is that candidates with financial resources to spend on superior LLMs may secure relatively more employment opportunities in organizations using verbal assessments, and the continued use of verbal assessments as selection tools may exacerbate social inequalities; that is, the rich will likely get richer. However, GPT-4 may eventually become free to use as well.

Overall, the performance of both LLMs on the quantitative test was very poor, with GPT-4 only slightly outperforming GPT-3.5.

Thus, it seems from this initial investigation that aptitude assessments involving numerical problems (particularly those that cannot easily be entered in Python via Code Interpreter) are more robust to candidates' use of LLMs as an aide. Interestingly, however, GPT-4's performance improved substantially when both the format of the quantitative assessment was adapted into a multi-choice test and the missing number in the sequence was positioned at the end. Likely, this occurs because LLMs are trained to work from left to right in language, and thus, reasoning that requires moving both left and right (such as is needed when the missing number is not at the end) may be more difficult for the LLMs. These results suggest that subtle decisions regarding test item design may have important consequences for test vulnerability to LLMs. To the extent that these results generalize to other types of tests involving sequences, test developers may be wise to focus on questions where the missing item in the sequence is not in the final spot. Nonetheless, future research is required to understand whether the results observed here generalize to other types of test items that involve sequences (e.g., Condon & Revelle, 2014) or if the pattern is only applicable to numerical sequences.

Although emerging research has demonstrated that different prompt approaches can elicit responses of varying quality from LLMs (White et al., 2023), our comparisons of six prompt approaches suggested that prompting plays a relatively minor role in determining performance on the quantitative and verbal ability tests. Excepting the bare bones prompt approach with GPT-3.5 on the verbal ability test, the largest differences, in raw scores, between the least and most effective prompt approaches did not go beyond 4 points (or 12.03 percentile improvement). Further, these "large" differences were only observed for the quantitative ability test, where the models struggled to perform well. The bare bones approach may have exhibited poor performance because it did not include instructions to think out the answer step-by-step, which as noted in Footnote 4, tended to improve LLM performance on the verbal ability (but not the quantitative ability) test. Additionally, without additional instructions, it may have considered the accuracy of the statement in relation to the broader world, rather than only in relation to the passage of text provided to it.

While recognizing that we only examined six strategies from an infinite set, these initial findings suggest that variation in candidates' adoption of prompt approaches is unlikely to lead to major differences in performance between candidates using LLMs—particularly with GPT-4. Indeed, as LLMs become more powerful, they better understand language, which is likely to reduce the importance of prompting (Acar, 2023). One potential implication for this result is that test developers will need to continuously monitor the psychometric properties and mean scores of their tests, as extensive LLM-based cheating may affect the distributions of test scores. For example, it may be that secondary (or additional) modes that represent the maximum test performance of a popular LLM may emerge over time, which may influence the transformation of raw test scores into percentiles.

The temperature setting of LLMs affects how much randomness exists in the model. While it may be reasonable to assume that a

lower temperature, by increasing the consistency of outputs, would give better answers, we found that a nonzero temperature tended to provide small performance improvements. Thus, it seems that having some randomness in generation may have helped the LLMs come to the correct response, whereas the less random generation when temperature equals zero made the LLMs less likely to uncover the correct answer. Overall, the default temperature settings for ChatGPT benefits applicants, relative to a more conservative temperature setting, although the benefit seems to be very small.

6.2 | Implications for unproctored cognitive ability testing

Even before the threat of candidates using LLMs to complete tests on their behalf, the use of cognitive ability testing for selection was recently called into question by academics due to concerns around the initial over-estimation of their criterion-related validity (Sackett et al., 2022; Sackett et al., 2024; cf., Schmidt & Hunter, 1998) and the adverse impact these tests can have on disadvantaged group members (e.g., Woods & Patterson, 2024). Our results suggest that LLMs have created new problems for unproctored testing, particularly for verbal ability tests, given that they can perform better than most human test takers and are widely available for people to use. In other words, tests that were assumed to be costly to fake may now become less trustworthy signals of a candidate's ability (Bangerter et al., 2012). This alters the testing landscape, because the continued adoption of unproctored cognitive ability testing may trigger a further degradation of observed criterion-related validity as more applicants use LLMs to generate responses. Further, socioeconomic inequality may be exacerbated as candidates with the resources to use the superior LLMs outperform those who cannot afford to. One possible outcome is that employers may adapt their selection systems to remove ability tests that come to be regarded as easy to fake. Alternatively, reintroducing proctoring could in turn reintroduce significant risks to candidates wishing to cheat, although it will also increase employer costs.

Nonetheless, while the performances of the LLMs on the verbal ability test was impressive, this was not the case for the quantitative test, where even GPT-4 performed worse than a large majority of the norm group. Thus, in practice, a candidate who relies on an LLM to complete both tests would likely not be among the most highly ranked. Further, we also recognize that, currently, there remain some practical constraints preventing candidates from being able to completely rely on an LLM to complete a test on their behalf. First, candidates must be able to transcribe the test materials into the LLM application. Preventing test-takers from copying and pasting item text is technically very straightforward (e.g., by delivering the item as an image instead of text). Second, many ability tests—including those studied here—are timed, limiting applicant capacity for relying on outside sources, and at the time of writing, the speed and dependability of LLMs were variable. For example, in our attempt at the bare-bones prompt using the ChatGPT website, we found that

GPT-4's Code Interpreter sometimes took 3 min to generate a response, and often that response would not even include a suggested solution. And third, the current version of ChatGPT places heavy restrictions on the number of messages allowed in a given period. Nonetheless, we suspect that, over time, these barriers and others will be easily overcome as LLM and supporting technologies evolve. For example, alleviating both the transcription and time problems, the current version of the ChatGPT cell phone app can integrate with a phone's camera and automatically transcribe text from a photo image. It is therefore vital to remain alert to the development in the capabilities of LLMs (see Landers, 2023).

6.3 | Limitations, future research, and conclusion

Our study is limited in several ways that suggest directions for future research. First, because we were examining multiple conditions in multiple trials across multiple tests, we prompted the LLMs in a way that encouraged them to provide their final answer in a format that was easily retrievable with computer code. Specifically, we asked the LLM to provide its final answer in curly brackets {}. However, in doing so, we may have altered the behavior of the LLM, given that any alterations to the input sequence can substantially alter results.

Second, to facilitate the investigation, we used code to automatically retrieve the final answer and gave the LLM no opportunities to improve upon its response to the initial prompt. However, applicants could iteratively query an LLM. For example, in a multiple-choice test, if the LLM returned no answers that matched the response options, a user could ask the LLM to reconsider and revise its output. Indeed, in our bare bones run on Code Interpreter with the quantitative test, the ChatGPT app would often request additional information from the user. This could potentially improve LLM test performance, but it remains to be seen how user interaction with LLMs influences LLM test performance. It is possible that applicants of higher levels of ability will be more effective at querying the LLM and, thus, can utilize LLMs more effectively than less able applicants.

Third, although we investigated a quantitative and verbal ability test, these were specific tests that each consisted of a single type of item. LLMs may perform better or worse on different item types, and there are many item types that we did not explore. For example, matrix reasoning items (e.g., Raven's standard progressive matrices) are commonly used, yet using an LLM on such items would require a multimodal LLM that can accept images as inputs. Initial evidence suggests that LLMs score well on matrix reasoning items when they are translated into text-based questions (Webb et al., 2023). Future work could investigate additional item types in validated tests used for high-stakes selection.

Fourth, we examined only a specific set of LLMs—namely, GPT-3.5 and GPT-4. Although these are extremely popular LLMs, many other LLMs exist and more are being created. For example, Alphabet (Google's parent company) just released Gemini, which was trained to

be multimodal. Future LLMs may perform better on quantitative ability tests, but current state-of-the-art LLMs can already complete verbal ability tests comparably to humans with extremely high verbal ability. Of particular interest here is the possibility that candidates had already used GPT 3.5 to cheat on the tests studied here, which would potentially put them into GPT-4's training data. A future study could examine whether open source LLMs perform better on test items in their training data than on items not in the training data. If so, test developers would need to be especially vigilant for items that are in publicly available, internet datasets.

Overall, our results surface fresh concerns regarding the use of unproctored cognitive ability testing for pre-employment assessment. The findings are particularly concerning for verbal ability tests, given that all prompt approaches with GPT-4 hold potential for nullifying the validity of such tests. However, if applicant quantitative ability is being tested, or if candidates use GPT-3.5, they may fare worse when using LLMs than when not using them.

DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available in the supplementary material of this article

ORCID

Patrick D. Dunlop  <http://orcid.org/0000-0002-5225-6409>

ENDNOTES

- 1 Notably, a version of GPT-4 is accessible via Microsoft Bing but the model is tweaked to also engage in real-time web search.
- 2 In our study, we developed prompt approaches, applied them, and examined whether the provided answer was correct. We did not iteratively prompt to improve outputs.
- 3 For interested readers, the OSF repository contains the code that shows how to specify the temperature setting when calling the API.
- 4 Given that this is a recommended step for improving the output of LLMs (e.g., <https://neurips.cc/virtual/2023/poster/71210>), we also explored this for the quantitative ability test. While this addition to the prompt improved performance with the verbal ability test, it tended to decrease performance on the quantitative ability test.
- 5 We also explored, via the ChatGPT website, using GPT-4 with the Code Interpreter feature and the bare bones prompt to respond to the quantitative ability test. The score was identical to using GPT-4 through the API without Code Interpreter. Code Interpreter enables users to upload files, run Python code, and generate files for download, when interacting with the LLM.

REFERENCES

- Acar, O. A. (2023). AI prompt engineering isn't the future. *Harvard Business Review*. Available at <https://hbr.org/2023/06/ai-prompt-engineering-isnt-the-future>
- Ali, R., Tang, O. Y., Connolly, I. D., Fridley, J. S., Shin, J. H., Zadnik Sullivan, P. L., Cielo, D., Oyelese, A. A., Doberstein, C. E., Telfeian, A. E., Gokaslan, Z. L., & Asaad, W. F. (2023). Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. *Neurosurgery*, 93(5), 1090–1098. <https://doi.org/10.1227/neu.0000000000002551>
- Anthropic. (2024). *Prompt engineering*. Anthropic. <https://docs.anthropic.com/claude/docs/prompt-engineering>

- Arctic Shores. (2023a). *ChatGPT vs Personality Assessments*. Arctic Shores. <https://www.arcticshores.com/insights/chatgpt-vs-personality-assessments-does-it-have-the-right-personality-traits-to-get-an-interview>
- Arctic Shores. (2023b). *ChatGPT vs Situational Judgement Tests: Can it outperform a human?* Arctic Shores. <https://www.arcticshores.com/insights/chatgpt-vs-situational-judgement-tests-how-it-performs-vs-a-human>
- Arthur, Jr., W., Glaze, R. M., Villado, A. J., & Taylor, J. E. (2010). The magnitude and extent of cheating and response distortion effects on unproctored Internet-based tests of cognitive ability and personality. *International Journal of Selection and Assessment*, 18(1), 1–16. <https://doi.org/10.1111/j.1468-2389.2010.00476.x>
- Bangerter, A., Roulin, N., & König, C. J. (2012). Personnel selection as a signaling game. *Journal of Applied Psychology*, 97(4), 719–738. <https://doi.org/10.1037/a0026078>
- Beaty, J. C., Nye, C. D., Borneman, M. J., Kantrowitz, T. M., Drasgow, F., & Grauer, E. (2011). Proctored versus unproctored internet tests: Are unproctored noncognitive tests as predictive of job performance? *International Journal of Selection and Assessment*, 19(1), 1–10. <https://doi.org/10.1111/j.1468-2389.2011.00529.x>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.
- Borchert, R. J., Hickman, C. R., Pepys, J., & Sadler, T. J. (2023). Performance of ChatGPT on the situational judgement test—A professional Dilemmas-based examination for doctors in the United Kingdom. *JMIR Medical Education*, 9, e48978. <https://doi.org/10.2196/48978>
- Brin, D., Sorin, V., Vaid, A., Soroush, A., Glicksberg, B. S., Charney, A. W., Nadkarni, G., & Klang, E. (2023). Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Scientific Reports*, 13(1), 16492. <https://doi.org/10.1038/s41598-023-43436-9>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Budhwar, P., Chowdhury, S., Wood, G., Aguinis, H., Bamber, G. J., Beltran, J. R., Boselie, P., Lee Cooke, F., Decker, S., DeNisi, A., Dey, P. K., Guest, D., Knoblich, A. J., Malik, A., Paauwe, J., Papagiannidis, S., Patel, C., Pereira, V., Ren, S., ... Varma, A. (2023). Human resource management in the age of generative artificial intelligence: Perspectives and research directions on ChatGPT. *Human Resource Management Journal*, 33, 606–659. <https://doi.org/10.1111/1748-8583.12524>
- Chamorro-Premuzic, T., & Furnham, A. (2005). *Personality and Intellectual Competence*. Lawrence Erlbaum Associates.
- Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2023). Unleashing the potential of prompt engineering in large language models: A comprehensive review. *arXiv*, 2310.14735. <https://doi.org/10.48550/arXiv.2310.14735>
- Condon, D. M., & Revelle, W. (2014). The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence*, 43, 52–64. <https://doi.org/10.1016/j.intell.2014.01.004>
- Donovan, J. J., Dwight, S. A., & Hurtz, G. M. (2003). An assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the randomized response technique. *Human Performance*, 16(1), 81–106. https://doi.org/10.1207/S15327043HUP1601_4
- Elyoseph, Z., Hadar-Shoval, D., Asraf, K., & Lvovsky, M. (2023). ChatGPT outperforms humans in emotional awareness evaluations. *Frontiers in Psychology*, 14, 1199058. <https://doi.org/10.3389/fpsyg.2023.1199058>
- Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4), 681–694. <https://doi.org/10.1007/s11023-020-09548-1>
- Groza, A. (2023). Measuring reasoning capabilities of ChatGPT. *arXiv*, 2310.05993. <https://arxiv.org/abs/2310.05993>
- Hu, J., & Connelly, B. S. (2021). Faking by actual applicants on personality tests: A meta-analysis of within-subjects studies. *International Journal of Selection and Assessment*, 29(3–4), 412–426. <https://doi.org/10.1111/ijasa.12338>
- Kantrowitz, T. M., & Dainis, A. M. (2014). How secure are unproctored pre-employment tests? Analysis of inconsistent test scores. *Journal of Business and Psychology*, 29(4), 605–616. <https://doi.org/10.1007/s10869-014-9365-6>
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2023). Large language models are zero-shot reasoners. *arXiv*, 2205.11916v4. <https://doi.org/10.48550/arXiv.2205.11916>
- Landers, R. N. (2023). Fixing the industrial-organizational psychology-technology interface (IOPTI): Avoiding both IO/Tech and Tech/IO conflict. In T. M. Kantrowitz, D. H. Reynolds, & J. C. Scott (Eds.), *Talent assessment: Embracing innovation and mitigating risk in the digital age* (pp. 202–218). Oxford University Press. <https://doi.org/10.1093/oso/9780197611050.003.0013>
- Levashina, J., & Campion, M. A. (2007). Measuring faking in the employment interview: Development and validation of an interview faking behavior scale. *Journal of Applied Psychology*, 92(6), 1638–1656. <https://doi.org/10.1037/0021-9010.92.6.1638>
- Levashina, J., Morgeson, F. P., & Campion, M. A. (2009). They don't do it often, but they do it well: Exploring the relationship between applicant mental abilities and faking. *International Journal of Selection and Assessment*, 17(3), 271–281. <https://doi.org/10.1111/j.1468-2389.2009.00469.x>
- Lievens, F., & Burke, E. (2011). Dealing with the threats inherent in unproctored Internet testing of cognitive ability: Results from a large-scale operational test program. *Journal of Occupational and Organizational Psychology*, 84(4), 817–824. <https://doi.org/10.1348/096317910X522672>
- Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. *arXiv*, 2005.00661. <https://doi.org/10.48550/arXiv.2005.00661>
- Metcalfe, C. (2024). Job titles of the future: AI prompt engineering. *MIT Technology Review*, May/June. Available at <https://www.technologyreview.com/2024/04/24/1091125/ai-prompt-engineer-generative-ai-job-titles/>
- Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences of the United States of America*, 120(13), 1–5. <https://doi.org/10.1073/pnas.2215907120>
- Mitchell, M., Palmarini, A. B., & Moskvichev, A. (2023). Comparing humans, GPT-4, and GPT-4V on abstraction and reasoning tasks. *arXiv*, 2311.09247v09242. <https://doi.org/10.48550/arXiv.2311.09247>
- Morton, J. (2024). Using prompt engineering to better communicate with people. *Harvard Business Review*. <https://hbr.org/2024/01/using-prompt-engineering-to-better-communicate-with-people>
- Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramer, F., & Lee, K. (2023). *Extracting training data from ChatGPT*. Available at <https://not-just-memorization.github.io/extracting-training-data-from-chatgpt.html>
- Nye, C. D., Do, B.-R., Drasgow, F., & Fine, S. (2008). Two-step testing in employee selection: Is score inflation a problem? *International Journal of Selection and Assessment*, 16(2), 112–120. <https://doi.org/10.1111/j.1468-2389.2008.00416.x>

- OpenAI. (2023). *GPT-4 Technical Report*. OpenAI. <https://cdn.openai.com/papers/gpt-4.pdf>
- Phillips, J., & Robie, C. (2024). Can a computer outfake a human? *Personality and Individual Differences*, 217, 112434. <https://doi.org/10.1016/j.paid.2023.112434>
- Porter, J. (2023). ChatGPT continues to be one of the fastest-growing services ever. *TheVerge*. <https://www.theverge.com/2023/11/6/23948386/chatgpt-active-user-count-openai-developer-conference>
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. Available at <https://www.mikecaptain.com/resources/pdf/GPT-1.pdf>
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 9.
- Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3, 121–154. <https://doi.org/10.1016/j.iotcps.2023.04.003>
- Sackett, P. R., Demeke, S., Bazian, I. M., Griebie, A. M., Priest, R., & Kuncel, N. R. (2024). A contemporary look at the relationship between general cognitive ability and job performance. *The Journal of applied psychology*, 109, 687–713. <https://doi.org/10.1037/apl0001159>
- Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2022). Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology*, 107(11), 2040–2068. <https://doi.org/10.1037/apl0000994>
- Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124(2), 262–274. <https://doi.org/10.1037/0033-2909.124.2.262>
- Schreiner, M. (2023). *GPT-4 architecture, datasets, costs, and more leaked*. Available at <https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/>
- Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. *arXiv*, 2303.11366. <https://doi.org/10.48550/arXiv.2303.11366>
- Similar Web. (2023). *Traffic analytics, ranking stats, and tech stack*. Similar Web. <https://www.similarweb.com/website/chat.openai.com/#demographics>
- Steger, D., Schroeders, U., & Gnams, T. (2020). A meta-analysis of test scores in proctored and unproctored ability assessments. *European Journal of Psychological Assessment*, 36, 174–184. <https://doi.org/10.1027/1015-5759/a000494>
- Templer, K. J., & Lange, S. R. (2008). Internet testing: Equivalence between proctored lab and unproctored field conditions. *Computers in Human Behavior*, 24(3), 1216–1228. <https://doi.org/10.1016/j.chb.2007.04.006>
- The International Cognitive Ability Resource Team (ICAR) (2014). <https://icar-project.com/>
- Tippins, N. T., Beaty, J., Drasgow, F., Gibson, W. M., Pearlman, K., Segall, D. O., & Shepherd, W. (2006). Unproctored internet testing in employment settings. *Personnel Psychology*, 59(1), 189–225. <https://doi.org/10.1111/j.1744-6570.2006.00909.x>
- Wang, B., Min, S., Deng, X., Shen, J., Wu, Y., Zettlemoyer, L., & Sun, H. (2022). Towards understanding chain-of-thought prompting: An empirical study of what matters. *arXiv*, 2212.10001v2. <https://doi.org/10.48550/arXiv.2212.10001>
- Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9), 1526–1541.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *arXiv*, arXiv:2201.11903v6. <https://doi.org/10.48550/arXiv.2201.11903>
- White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv*, rXiv:2302.11382v1. <https://doi.org/10.48550/arXiv.2302.11382>
- Woods, S. A., & Patterson, F. (2024). A critical review of the use of cognitive ability testing for selection into graduate and higher professional occupations. *Journal of Occupational and Organizational Psychology*, 97(1), 253–272. <https://doi.org/10.1111/joop.12470>
- Xu, B., Yang, A., Lin, J., Wang, Q., Zhou, C., Zhang, Y., & Mao, Z. (2023). Expert prompting: Instructing large language models to be distinguished experts. *arXiv*, 2305.14688. <https://doi.org/10.48550/arXiv.2305.14688>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Hickman, L., Dunlop, P. D., & Wolf, J. L. (2024). The performance of large language models on quantitative and verbal ability tests: Initial evidence and implications for unproctored high-stakes testing. *International Journal of Selection and Assessment*, 32, 499–511. <https://doi.org/10.1111/ijsa.12479>