Is Machine Psychology here? On Requirements for Using Human Psychological Tests on Large Language Models

Lea Löhn*, **Niklas Kiehne***, **Alexander Ljapunov, Wolf-Tilo Balke**
Institute for Information Systems
TU Braunschweig
Braunschweig, Lower Saxony, Germany

Abstract

In an effort to better understand the behavior of large language models (LLM), researchers recently turned to conducting psychological assessments on them. Several studies diagnose various psychological concepts in LLMs, such as psychopathological symptoms, personality traits, and intellectual functioning, aiming to unravel their black-box characteristics. But can we safely assess LLMs with tests that were originally designed for humans? The psychology domain looks back on decades of developing standards of appropriate testing procedures to ensure reliable and valid measures. We argue that analogous standardization processes are required for LLM assessments, given their differential functioning as compared to humans. In this paper, we propose seven requirements necessary for testing LLMs. Based on these, we critically reflect a sample of 25 recent *machine psychology* studies. Our analysis reveals (1) the lack of appropriate methods to assess test reliability and construct validity, (2) the unknown strength of construct-irrelevant influences, such as the contamination of pre-training corpora with test material, and (3) the pervasive issue of non-reproducibility of many studies. The results underscore the lack of a general methodology for the implementation of psychological assessments of LLMs and the need to redefine psychological constructs specifically for large language models rather than adopting them from human psychology.

1 Introduction

Large language models (LLM) demonstrate surprisingly strong natural language generation abilities across a range of tasks (Srivastava et al., 2023), sparking debates about the emergence of human characteristics, such as personality traits, empathy, intuitive reasoning, ethical understanding, or even traits of sentience, see e.g. (Miotto et al., 2022; Kosinski, 2023; Hagendorff et al., 2022; Kiehne et al., 2024; Blum and Blum, 2024). Yet recently, experts raised concerns about their inherent opaqueness and the potential dangers that could follow their widespread adoption (Dale, 2021; Future of Life Institute, 2023). This incomprehensibility of the inner workings and decision processes of current LLMs prompted researchers to borrow methods from human psychology to shed light on the behavior of these black-box models: LLMs are analyzed via psychological assessments, often referred to as *machine psychology* or *AI psychometrics* (Hagendorff, 2023; Pellert et al., 2024). Kosinski (2023) utilizes an unexpected contents task to diagnose Theory of Mind in GPT-4, arguing that the ability to ascribe mental states emerges with sufficient model size. Yet, Ullman (2023) shows that trivial changes to the test items lead to the opposite outcome implying that GPT-4 does *not* have Theory of Mind. Arguably, these contrary results are symptomatic of a general lack of standardization in the domain. Meanwhile, the number of machine psychology studies grows quickly across various psychological constructs. The aim of this paper is to provide a solid foundation for using psychological tests on LLMs. Indeed, many studies haphazardly use psychological tests without properly incorporating necessary theoretical underpinnings. Grounded in the well-established standards of psychological testing, we propose seven essential requirements for test use in machine psychology. We thus advocate for stricter rules governing reliable, valid, and fair testing, also taking into consideration the quirks of current LLMs, such as their sensitivity to wording. As a proof of concept, we critically reflect 25 recent works regarding these requirements, highlighting the unresolved issues in the field. Our analysis clearly challenges the evidential and declarative power of current methodologies for the psychological assessments of LLMs, while also providing a more reliable foundation.

*Correspondance to:
{lealoehn, niklas.kiehne}@gmail.com

Proceedings of the 17th International Natural Language Generation Conference, pages 230–242
September 23–27, 2024.

2 Background

The assessment of humans on diverse psychological constructs has been at the core of psychology as a scientific domain, dating back to at least the 19th century (Galton, 1869). The term *construct* refers to a group of psychological characteristics, such as behavioral patterns, personality traits or cognitive skills (Slaney and Garcia, 2015). A construct is often defined conceptually by abstractly describing its meaning and relations to other constructs, and operationally by stating variables used to measure it (Reynolds and Livingston, 2019).

The methods of psychological testing have been the subject of rigorous research for decades, aiming to enhance their reliability, validity, and overall effectiveness in assessing various aspects of human cognition, personality, and behavior (American Educational Research Association et al., 2014; Reynolds and Livingston, 2019). An important aspect in this regard is the formalization and standardization of correct assessment practices, concerning test development, application, and evaluation. We consider two widely accepted standards, namely the *Standards for Educational and Psychological Testing* and the *International Guidelines for Test Use* (American Educational Research Association et al., 2014; International Test Commission, 2001), henceforth referred to as the *Standards*. These guidelines are designed for test developers, administrators, and users to promote best practices and ethical standards in psychological testing.

In contrast, the comparatively young machine psychology domain has not yet settled on such standards. The field itself is still developing, often using different terminology. Rahwan et al. (2019) propose the broad term *machine behavior* to combine methods from various sciences to better understand AI agents. Pellert et al. (2024) suggest the area of *AI psychometrics* as a combination of psychology, computer science and linguistics. Similarly, Hagendorff (2023) introduce machine psychology as an umbrella term, which we will adopt throughout the paper. Pellert et al. (2024) and Hagendorff (2023) argue that machine psychology differs from LLM benchmarking by focusing on diagnosis rather than establishing performance.

Some efforts towards a more standardized methodology have been made by Hagendorff (2023), who proposes a set of guidelines that should be considered when conducting machine psychology studies. These guidelines mainly focus on prompt design, given the significant impact it has on prompt completions of generative language models. Frank (2023) suggests a combination of methods from developmentalists and computational scientists that could assist in uncovering abstract representations in language models. However, both of these approaches focus on the practical design of machine psychology studies rather than the question of which general criteria should be taken into account for a psychological test to be a meaningful assessment tool for LLMs. Our work aims to provide a first set of normative requirements, prioritizing strictly necessary pre-requisites of correct testing over technical possibility. We build on proven methodologies from traditional psychology and show their applicability to the machine psychology domain.

3 Requirements for Machine Psychology

We extracted and summarized a list of requirements for psychological testing from the *Standards* that play a pivotal role in the selection, administration and scoring of tests. When conducting psychological assessments of LLMs, certain requirements that apply to human psychological assessments may not be necessary to consider. All requirements concerning the test taker's data privacy, for example, are inapplicable when the test subject is a machine. Due to the different characteristics of human and AI examinees, the resulting requirements were transferred and adapted to the AI domain. We want to emphasize that our requirements are derived solely from psychological testing theory. This is important because the correct application of the tests used in machine psychology primarily depends on psychological standards, rather than LLM evaluation practices. Nonetheless, we find some of our derived requirements to have well-known counterparts in the general LLM evaluation domain. For example, the contamination of pre-training corpora with benchmarking material is a fundamental problem affecting virtually all evaluation methodologies, including those of machine psychology (Jacovi et al., 2023; Sainz et al., 2023). The proposed list of requirements is not intended to be exhaustive, but instead provides an important basis of prerequisites to consider. We argue that these requirements should be fulfilled in the assessment of LLMs in order to provide meaningful results.

3.1 (R1) Reliability for the Intended Use

Reliability refers to the stability of test scores over multiple runs of the test. It can be affected by any kind of variability during repetitions of the testing procedure that can occur either as a result of factors internal to the test taker (e.g. motivation, attention or interest) or externally as a consequence of testing conditions and scoring procedure. The reliability of test scores may vary depending on the population under consideration, as the impact of those different variabilities in the testing process can differ for populations (American Educational Research Association et al., 2014).

Language models are influenced by many factors, e.g. architecture, training data, and hyperparameters, among others. Thus, it is evident that test reliability is not guaranteed across different models and that it must be carefully addressed. As a general principle, test reliability must be ensured for each considered population separately, including LLMs. Popular measures, such as test-retest, alternate-forms, or the internal consistency method, work independently of the nature of test takers and could be readily applied in the LLM domain. Interestingly, high test-retest reliability can be achieved by reducing the influence of randomness in the generation procedure, e.g. by lowering the temperature during sampling. In fact, deterministic generation modes can even guarantee perfect test-retest reliability, although these setups cover only a small fragment of the behaviors and thus can not accurately represent the full model. More importantly, simple test repetition does not suffice to account for model specific phenomena, such as their unusual sensitivity to input variations (Kiehne et al., 2024; Elazar et al., 2021). Here, multiple rephrased tests (alternate-forms) or a comparison of test items that measure the same component of a construct (internal consistency) are needed.

3.2 (R2) Validity for the Intended Use

The most important requirement for psychological tests is that the interpretation of test results is backed by theoretical frameworks and empirical evidence, a characteristic generally referred to as *validity*. In other words, it must be proven that a test indeed measures the construct it is intended to measure. Validity evidence can be provided based on the test content, the response processes of the test takers, the internal structure of the test and the relations to other variables (American Educational Research Association et al., 2014).

Evidence based on test content is obtained by analyzing the relationship between a test's content (e.g., themes, format, and wording) and the construct to be measured. It is important to examine how well the content domain is represented by the chosen test content and evaluate its relevance to the intended interpretations. This is often done by expert judges. Evidence based on response processes can be obtained by analyzing the degree to which the cognitive processes and strategies test takers use while responding to test items are in accordance with the intended construct. The analysis is usually done by performing interviews with different groups of test takers about their response strategies, but, depending on the construct measured, can also include investigations of physiological variables, such as eye-movement. Evidence based on internal structure evaluates how well the relationships between test items align with the proposed construct. An analysis should determine whether a hypothesized multidimensional construct is reflected in the test's internal structure. This is often done using factor analysis, which identifies the distinct factors the test is based on. Evidence based on relations to other variables can be provided by analyzing the relationship of test scores with external variables. This includes assessing the relationships to different tests that measure the same or associated constructs (convergent evidence) or relations to tests purportedly assessing different constructs (discriminant evidence).

3.3 (R3) Suitability for Test Takers

Any psychological assessment has to account for the capabilities and characteristics of its test takers, including, but not limited to, their cognitive abilities and sensory perceptions (American Educational Research Association et al., 2014). Similar arguments hold for language models. Here, it is required that tests fit the supported in- and output formats. For example, generative language models should only be exposed to written tests requiring textual answers, whereas a text-classification system is unable to produce free-form text.

3.4 (R4) Non-Disclosure of Test Materials

In psychological assessments of humans, it is crucial to ensure that examinees have not been exposed to the test material prior to the assessment in order to avoid biased and invalid results (American Educational Research Association et al., 2014).

Similar biases have been observed in the generated responses of LLMs, as they have been shown to perfectly replicate patterns from their training data (Nasr et al., 2023; Emami et al., 2020). Thus, in our context, the requirement translates to ensuring that the training data of the models does not contain any test material. Naturally, the question arises whether the massive pre-training corpora of contemporary state-of-the-art models are in fact contaminated by test material, and also, to which extent this effect impacts the testing process. Emami et al. (2020) show that the overlap of testing and training data significantly affects model performance, suggesting that if contamination occurred, then it will likely re-emerge during testing. Therefore, researchers must either show the absence of these effects on original tests or take measures to ensure the uniqueness of the test material.

3.5 (R5) Fairness

The central idea of fairness in testing is to minimize construct-irrelevant influences on test score variance and thus, to support comparable interpretations across all examinees.

(R5a) Test Validity for all Models It is a common practice to compare different language models regarding their performance on various benchmarks. Similarly, researchers in the field of machine psychology seek to compare the psychological characteristics of several LLMs. In such comparative studies, it is of critical importance to ensure that the results being compared were obtained from a test that has been validated for all models being considered for comparison.

(R5b) Validity of Test Translations Many generative language models can be operated in different languages, thereby allowing the psychological testing of models in a range of languages. When choosing the test language, it is important to not only consider the test taker's proficiency in that language, but also to ensure that the translation is validated. A multitude of psychological tests have already been validated in different languages, with published versions available. When translating independently, it is advisable to adhere to established conventions, such as those set out in (International Test Commission, 2017).

(R5c) Transparent Test Use Similarly to tests conducted on humans, machine psychology tests need standardized and transparent evaluation procedures to allow for valid comparisons and interpretations. The generation process of many LLMs can be controlled via a multitude of sampling methods and parameters, often referred to as decoding strategies (Holtzman et al., 2020). Das and Balke (2022) show that each component in the decoding process might impact how biases are propagated into the generated responses. Thus, test scores may vary significantly for the same LLM, depending on the exact evaluation procedure. The wording of instructions and test items can also have strong impacts on model behavior. These manifold influences on test scores call for researchers to prioritize transparent and reproducible test use to allow for comparability between multiple studies. This includes the complete testing setup, e.g., model weights, in- and output formatting, and parameters of the generation process.

4 Analysis of Machine Psychology Studies

In this section, we analyze various studies in the machine psychology domain concerning the requirements R1-R5c identified in Section 3. The initial pool of literature was collected up until October 2023 using keyword searches on popular databases, such as Google Scholar[1], Scopus[2], and DBLP[3]. After title and abstract screening, we traced the citation network[4] to further augment the literature pool. We retain 25 papers in which researchers investigated a total of 12 different psychological constructs using 34 different psychological tests and assessments. A detailed analysis of the application areas is presented in Tables 2 and 3 in the appendix. As the machine psychology domain is currently emerging, the studies we considered are rather recent, with publication dates ranging from June 2022 to September 2023. The domain enjoys research contributions from scholars of diverse fields, ranging from psychology, social sciences, economics, cognitive, and computer sciences.

4.1 Overview of the Literature

Most studies aim to assess the cognitive functions and personality traits of LLMs. Others investigate Theory of Mind (Bubeck et al., 2023; Kosinski, 2023; Trott et al., 2023; Ullman, 2023), creativity (Goes et al., 2023; Haase and Hanel, 2023; Stevenson et al., 2022; Summers-Stay et al.,

[1] https://scholar.google.com
[2] https://www.scopus.com
[3] https://dblp.org
[4] https://www.connectedpapers.com

Paper	Reliability R1	Validity R2	Suitability R3	Non-Disclosure R4	Test Validity for all Models R5a	Validity of Test Translations R5b	Transparent Test Use R5c
Aher et al. (2023)	❍	❍	●	●	✗	–	●
Argyle et al. (2023)	✗	✗	●	✗	–	–	●
Binz and Schulz (2023)	●	❍	●	❍	–	–	●
Bubeck et al. (2023)	✗	✗	●	●	✗	–	✗
Chen et al. (2023)	❍	❍	●	✗	–	–	❍
Coda-Forno et al. (2023)	●	●	●	●	–	–	●
Dasgupta et al. (2022)	❍	❍	●	●	–	–	❍
Fischer et al. (2023)	✗	●	●	❍	–	–	●
Fraser et al. (2022)	●	❍	●	●	–	–	●
Goes et al. (2023)	✗	✗	●	✗	–	–	●
Haase and Hanel (2023)	●	❍	●	✗	✗	–	●
Hagendorff et al. (2023)	❍	❍	●	●	–	–	●
Horton (2023)	❍	❍	●	❍	✗	–	●
Jones and Steinhardt (2022)	✗	✗	●	✗	–	–	●
Kosinski (2023)	✗	●	●	●	✗	–	●
Li et al. (2023)	●	❍	●	✗	✗	–	●
Miotto et al. (2022)	●	❍	●	❍	–	–	●
Park et al. (2023)	●	●	●	❍	–	–	●
Pellert et al. (2024)	❍	❍	●	✗	✗	●	●
Serapio-García et al. (2023)	●	●	●	✗	●	–	●
Song et al. (2023)	●	❍	●	✗	❍	–	●
Stevenson et al. (2022)	●	❍	●	❍	–	❍	●
Summers-Stay et al. (2023)	❍	✗	●	❍	–	–	✗
Trott et al. (2023)	●	●	●	●	●	–	●
Ullman (2023)	●	●	●	●	–	–	●

Table 1: Assessment of requirements R1-R5c in 25 machine psychology studies. We denote requirements as: – not applicable, ✗ not addressed, ❍ discussed, ● appropriate effort/study conducted, but missing supporting evidence, ● any evidence of fulfillment provided. The symbols and the annotation process are explained in detail in Section 4.2.

2023), reasoning (Binz and Schulz, 2023; Chen et al., 2023; Hagendorff et al., 2023), and decision-making (Binz and Schulz, 2023; Chen et al., 2023; Horton, 2023; Park et al., 2023). Personality traits of LLMs are studied via classical personality tests (Li et al., 2023; Miotto et al., 2022), tests for dark personality traits (Li et al., 2023), personal value inventories (Fischer et al., 2023; Miotto et al., 2022), and their moral attitudes (Fraser et al., 2022). There are also studies regarding problem and adaptive behavior (Coda-Forno et al., 2023; Li et al., 2023). Models of the GPT family are among the most frequently studied, possibly due to their widespread popularity. In total, 20 of the 25 studies include GPT-3 or newer versions, out of which 17 do not consider any other model. The remaining LLMs include BLOOM (Scao et al., 2023), FLAN-PaLM (Chung et al., 2024), DELPHI (Jiang et al., 2021), BERT-derivatives (Devlin et al., 2019; Liu et al., 2019; Sanh et al., 2020), and Chinchilla (Hoffmann et al., 2022). Most of the analyzed studies have simply been performed for analysis and possible diagnosis of specific traits. In addition, the test results are usually compared with human norms or between different LLMs. Further studies include the manipulation of test results by inducing construct-related linguistic input to test prompts (Coda-Forno et al., 2023; Serapio-García et al., 2023), the priming of models with demographic information in order to simulate different

human participants (Aher et al., 2023; Argyle et al., 2023), and the analysis of instruction fine-tuning as a method to impact test results (Li et al., 2023).

4.2 Assessing Machine Psychology Requirements

Table 1 illustrates each studies' placement regarding our selected requirements. We (the authors) examined and evaluated the treatment of each requirement in the chosen studies in joint meetings, where we collectively decided on a ranking. In certain instances, a requirement was not applicable to all studies. This is the case for R5a when only one model was tested and for R5b when only one language was assessed. We indicate such instances as –. If a requirement or a problem associated with the non-fulfillment of the requirement was not mentioned at all, we assigned an ✗. Should the necessity for fulfillment of a requirement be identified, yet no action be taken, a ❍ was assigned. This may be the case if a requirement was discussed, e.g. as a limitation of the study or as suggestion for future work. In certain instances, efforts were made (for requirements R3, R4, R5b, R5c) and/or studies were conducted (for requirements R1, R2, R4, R5a, R5b) with the objective of fulfilling the requirements. An effort that lacks supporting evidence that the requirement has been fulfilled or a study that shows a non-fulfillment of a requirement is designated as ●. If any evidence of fulfillment is provided, we assign a ●. For requirements R1 (reliability) and R2 (validity) we assign ● if at least one investigation of reliability or validity as discussed in Section 3.1 and 3.2 was conducted. This would apply to R1, for example, if test executions were analyzed in different formulations, which accounts as a method to assess alternate forms reliability. In the same way we rate with ● if at least one of the possible studies has led to evidence of fulfillment. We want to emphasize that such a classification for R1 or R2 only acknowledges evidence of fulfillment of one form of reliability or validity, and thus, does not necessarily imply that full evidence of reliability or validity of the chosen test was provided.

R1 (Reliability for the Intended Use) In terms of the investigation of reliability, the studies analyzed have addressed different forms of this requirement. In tests that require a subjective judgment of the answers given by test takers, researchers consider interrater-reliability (Haase and Hanel, 2023; Stevenson et al., 2022). Other studies are able to provide evidence of internal consistency for the used tests by computing inter-facet correlations or applying common measures, such as Cronbach's Alpha (Miotto et al., 2022; Serapio-García et al., 2023). One of the key issues discussed in terms of reliability is the impact of different wordings of test items on test results, which can be seen as an investigation of alternate-forms reliability (Aher et al., 2023; Coda-Forno et al., 2023; Fraser et al., 2022). Similarly, Song et al. (2023) propose to demand *option-order symmetry* as a reliability criterion for scale-based personality tests. This criterion requires that a model chooses the same answer from a scale of answer options, regardless if given in ascending and descending order. They diagnose their tested models as not giving reliable answers because either option-order symmetry was violated, or the model always chose the same answer option regardless of semantics. The effects of different orders of answer options in multiple-choice questions are also investigated in other studies (Binz and Schulz, 2023; Coda-Forno et al., 2023; Park et al., 2023). Interestingly, Coda-Forno et al. (2023) are the only ones to derive evidence of reliability from their investigations of different orders of answer options.

R2 (Validity for the Intended Use) The majority of studies does not provide evidence of validity concerning the intended use, regardless of its enormous importance for the interpretation of test results. Coda-Forno et al. (2023) investigate the impact of anxiety test results on cognitive tasks, which is a form of convergent validity. It is important to note that in this approach, the test utilized as a comparison baseline was not validated for the use with large language models, making this method not strictly appropriate. Serapio-García et al. (2023) present the most comprehensive approach in this context: They define validity for LLM-based tests as observing conformity of test results and behavior in other tasks. Their validity study, consistent with psychological test development, examines reliability and various sources of validity, including convergent validity based on the correlation of personality test results with personality traits analyzed from generated texts by a psychologically validated tool. Fischer et al. (2023) change the original scale-based evaluation of the *Portrait Values Questionnaire* to a dictionary based approach for their assessment of ChatGPT. They

make use of an existing theory-driven value dictionary and perform an extensive validity study on the proposed evaluation procedure.

R3 (Suitability for Test Takers) The requirement for the suitability of tests for LLMs (R3) is the most addressed concern across all studies. This is due to the fact that we consider the utilization of a test with a suitable input and output format to be an appropriate effort. The requirement is considered fulfilled if the selected test has been originally designed in an appropriate format. Exceptions to this are, for example, the investigation of Theory of Mind. The original test requires children to be presented with specific scenarios, including dolls and objects, followed by questioning (Perner et al., 1987). Here, experiments of this sort are often transformed into text-based tests (Bubeck et al., 2023; Kosinski, 2023; Goes et al., 2023). Adaptions of the test material or the assessment itself, however, require new evidence of their validity in order to fulfill the requirement. No such evidence was found in the analyzed studies, resulting in a rating of ●.

R4 (Non Disclosure of Test Materials) Requirement R4 divides the literature into two camps: The majority of the studies do not take any measures to prevent the contamination of training data with test material. Some of these studies do, however, acknowledge this as a potential problem regarding the significance of test results. A common problem that researchers face in ruling out these effects is that the pre-training data is often not freely accessible. Unfortunately, especially the proprietary models, which currently enjoy the most interest by researchers and users, rarely allow access to their training datasets. Consequently, this requirement is often disregarded by researchers regardless of its high potential for skewing the test results (Emami et al., 2020). Nine out of 25 studies opt to modify the original test as a possible countermeasure. In this case, authors either rephrased items or generated entirely new test stimuli. Although modified tests may reduce the probability of LLMs having seen items before testing, evidence that such changes are still valid for the intended use is required. We acknowledged such evidence in only one study: Coda-Forno et al. (2023) compare the answers on rephrased and original test items and find a significant correlation, as well as no significant difference in the final test score.

R5a (Test Validity for all Models) When assessing multiple LLMs, requirement R5a demands proof of validity for each tested model. Out of the affected ten studies, only a single provides a thorough analysis in this regard (Serapio-García et al., 2023). The study underscores the importance of investigating the validity for all tested LLMs, as the authors conclude that larger, instruction-tuned models reach better results in the construct validity study.

R5b (Validity of Test Translations) With only two reports taking into account multilingual scenarios, requirement R5b is the least explored aspect among the specified requirements. Stevenson et al. (2022) include a translation of test answers to compare test scores of English and Dutch versions, which were separately administered to GPT-3 and a Dutch human group. The translation procedure was not further specified and as such, the comparability of both tests is hard to verify. In contrast, Pellert et al. only apply already validated translations (Pellert et al., 2024).

R5c (Transparent Test Use) While most of the studies make reasonable efforts to fulfill the requirement of a transparent testing procedure, only two out of 25 studies fully satisfy it. This is due to the fact that, although numerous studies publish model parameters or even code, they investigate proprietary models for which there is no guarantee that the version used will continue to be available in the future. This issue has a significant impact on their comparability and reproducibility.

5 Open Problems in Machine Psychology

Our analysis in the previous section demonstrates that there is no consensus among the selected papers regarding the requirements to be met in machine psychology studies. Moreover, not a single of the studies provides evidence of fulfillment of all requirements. Our assessments are also quite lenient, as we assign the highest possible grade whenever *any* evidence of fulfillment is presented. We intentionally did not rate the sufficiency of the evidence, as such judgments should be part of a broader scientific discourse.

The fundamental question when psychologically assessing large language models is whether a test validated as a measure of a specific construct for humans can also be a valid measure of that same construct for LLMs. This question remains unan-

swered in many studies of machine psychology. On closer examination the question opens up a number of problems, as discussed in the following.

Distinct Constructs for LLMs LLMs differ fundamentally from humans in their internal operations and external representations. Unlike humans, they lack a physical body to express any physiological variables of a construct measurement and operate only conditioned on their input, limiting their ability to experience the variety of situations that humans encounter in their daily life. This leads to the argument that construct definitions for humans might not be transferable to LLMs. Two issues follow.

First, comparisons of test scores for differing constructs might not be meaningful. In this case, comparing humans and LLMs could be potentially harmful and support misleading conclusions. Consequently, although still a common practice, it is currently inadvisable to compare the test results of humans and LLMs. Second, the contents of psychological tests might not be appropriate to measure the respective LLM construct. One solution could be to develop standalone construct definitions and corresponding tests for LLMs.

Unknown Response Processes The assumption underlying the administration of psychological tests is that the responses provided by test takers are the result of specific processes that align with the construct of interest. These cognitive processes and strategies are challenging to investigate for both human and LLM test takers, and can at best be approximated. Consequently, it remains unclear whether the internal response processes of humans and large language models are comparable at all, which makes the use of methods designed to isolate, trigger, and analyze human cognitive processes potentially unsuitable for large language models.

Validity of Modifications The current approaches to address reliability (R1), suitability (R3) and non-disclosure of test materials (R4) heavily rely on modifications of the original test items. Reliability is often measured by comparing the original test to variants of it, i.e. in a parallel forms setting, which the authors often derive themselves. To account for the in- and output modalities of their artificial test subjects, authors adapted original tests, e.g. by expressing interactive experiments in text-based stories. And finally, to evade the problem of training data contamination with test material, several papers chose to rephrase tests. Any modification of test items requires a re-validation of the changed material including empirical or logical evidence.

Individual or Population? One important difference in human and machine psychology is that the terms *individual* and *population* carry different meanings for LLMs and humans. From a psychological perspective, individuality requires self-awareness, autonomy, and agency, among others, and generally pertains to selfhood (Leary and Tangney, 2011). However, these three concepts alone are highly contentious in the general AI domain, as the scientific community has yet to reach consensus on whether they are at all achievable or even whether it is desirable to do so (Tegmark, 2018). Thus, although it is common practice to distinguish language models, e.g. by the configuration of their parameters, and consequently to refer to specific instances as *individuals*, it is advisable not to conflate this notion with those common in psychology.

In the analyzed studies, researchers have equated single LLMs both with an individual and with a population. However, many current LLMs can not guarantee stable and robust output behavior across multiple prompts and might even produce contradictory answers (Elazar et al., 2021; Kiehne et al., 2024). This stochastic nature of contemporary systems coupled with the fact that they incorporate data from oftentimes millions of different humans which has been shown to sporadically re-surface during answer generation, cast significant doubt on their qualification as individuals. While it is possible to extract meaningful population-level statistics from massively pre-trained models (Chu et al., 2023), this approach can not enumerate or even distinguish the individuals that the population comprises of. Additionally, Park et al. (2023) find an LLM's response distribution to be similar to that of a human population on some test items, but on others the model responds only with a singular answer – a pattern more akin to individuals. It remains unclear whether an LLM can truly be understood as an individual, which makes it tough to nail down to *what* exactly a test should apply. Currently, psychological tests on individuals do not find well-suited targets in the language model space.

6 Conclusion

We proposed a set of requirements that should be fulfilled for psychological assessments of large language models. These requirements were extracted from psychological standards and transferred to the LLM domain, asking for concrete actions to be taken. We then analyzed the extent to which our proposed requirements are currently being considered in a subsequent analysis of 25 studies from the machine psychology literature. Our findings reveal the lack of standardized testing procedures in the analyzed studies and clearly illustrate that the studies under review were not able to fulfill all of the requirements. Based on our investigations, we then derived a number of open problems in the field that show the current limitations of psychological assessments of LLMs. Our work contributes to this rapidly growing field of research by demonstrating the importance of standardized testing processes and providing a first framework of requirements to be considered in future works.

We want to stress that the requirements proposed in this paper can only scratch the surface of the vast theoretical landscape established in traditional psychology. Our work is limited in this regard. Further cooperative and interdisciplinary efforts are necessary to converge on a widely accepted standardization for the machine psychology domain. We hope this work encourages future studies to systematically address their results within the broader test-theoretical frameworks of psychology.

References

Gati V Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, pages 337–371, Honolulu, Hawaii, USA. JMLR.org.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. *Standards for educational and psychological testing.* American Educational Research Association, Washington, DC, US.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Lenore Blum and Manuel Blum. 2024. AI consciousness is inevitable: A theoretical computer science perspective. *Computing Research Repository*, arXiv:2403.17101. Version 3.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *Computing Research Repository*, arXiv:2303.12712. Version 5.

Yang Chen, Meena Andiappan, Tracy Jenkin, and Anton Ovchinnikov. 2023. A manager and an ai walk into a bar: Does chatgpt make biased decisions like we do? *SSRN Electronic Journal*.

Eric Chu, Jacob Andreas, Stephen Ansolabehere, and Deb Roy. 2023. Language models trained on media diets can predict public opinion. *Computing Research Repository*, arXiv:2303.16779. Version 1.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Julian Coda-Forno, Kristin Witte, Akshay K. Jagadish, Marcel Binz, Zeynep Akata, and Eric Schulz. 2023. Inducing anxiety in large language models increases exploration and bias. *Computing Research Repository*, arXiv:2304.11111. Version 1.

Robert Dale. 2021. Gpt-3: What's it good for? *Natural Language Engineering*, 27(1):113–118.

Mayukh Das and Wolf Tilo Balke. 2022. Quantifying bias from decoding techniques in natural language generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1311–1323, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ishita Dasgupta, Andrew K. Lampinen, Stephanie C. Y. Chan, Antonia Creswell, Dharshan Kumaran, James L. McClelland, and Felix Hill. 2022. Language models show human-like content effects on reasoning. *Computing Research Repository*, arXiv:2207.07051v1. Version 1.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving

consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Ali Emami, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2020. An analysis of dataset overlap on Winograd-style tasks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5855–5865, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ronald Fischer, Markus Luczak-Rösch, and Johannes A. Karl. 2023. What does chatgpt return about human values? exploring value bias in chatgpt using a descriptive value theory. *Computing Research Repository*, arXiv:2304.03612. Version 1.

Michael C Frank. 2023. Baby steps in evaluating the capacities of large language models. *Nature Reviews Psychology*, 2(8):451–452.

Kathleen C. Fraser, Svetlana Kiritchenko, and Esma Balkir. 2022. Does moral code have a moral code? probing delphi's moral philosophy. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, pages 26–42, Seattle, U.S.A. Association for Computational Linguistics.

Future of Life Institute. 2023. Pause giant ai experiments: An open letter. Accessed: 2023-11-23.

Francis Galton. 1869. *Hereditary Genius*. Macmillan and Co., London, Great Britain.

Fabricio Goes, Marco Volpe, Piotr Sawicki, Marek Grześ, and Jacob Watson. 2023. Pushing gpt's creativity to its limits: alternative uses and torrance tests. In *14th International Conference for Computational Creativity*.

Jennifer Haase and Paul H.P. Hanel. 2023. Artificial muses: Generative artificial intelligence chatbots have risen to human-level creativity. *Journal of Creativity*, 33(3):100066.

Thilo Hagendorff. 2023. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *Computing Research Repository*, arXiv:2303.13988. Version 4.

Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2022. Machine intuition: Uncovering human-like intuitive decision-making in GPT-3.5. *Computing Research Repository*, arXiv:2212.05206v1. Version 1.

Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nature Computational Science*, 3(10):833–838.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, et al. 2022. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, pages 30016–30030, New Orleans, LA, USA,. Curran Associates Inc.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia.

John J Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? Working Paper 31122, National Bureau of Economic Research.

International Test Commission. 2001. International guidelines for test use. *International Journal of Testing*, 1(2):93–114.

International Test Commission. 2017. The ITC guidelines for translating and adapting tests (second edition). `www.InTestCom.org`.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5084, Singapore. Association for Computational Linguistics.

Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jonathan Borchardt, et al. 2021. Delphi: Towards machine ethics and norms. *Computing Research Repository*, arXiv:2110.07574v1. Version 1.

Erik Jones and Jacob Steinhardt. 2022. Capturing failures of large language models via human cognitive biases. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, pages 11785–11799, New Orleans, LA, USA. Curran Associates, Inc.

Niklas Kiehne, Alexander Ljapunov, Marc Bätje, and Wolf-Tilo Balke. 2024. Analyzing effects of learning downstream tasks on moral bias in large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 904–923, Torino, Italia. ELRA and ICCL.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *Computing Research Repository*, arXiv:2302.02083v3. Version 3.

Mark R Leary and June Price Tangney. 2011. *Handbook of self and identity*. Guilford Press, New York, NY, US.

Xingxuan Li, Yutong Li, Linlin Liu, Lidong Bing, and Shafiq R. Joty. 2023. Does gpt-3 demonstrate psychopathy? evaluating large language models from a psychological perspective. *Computing Research Repository*, arXiv:2212.10529v2. Version 2.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, et al. 2019. Roberta: A robustly optimized BERT pretraining approach. *Computing Research Repository*, arXiv:1907.11692. Version 1.

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

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, et al. 2023. Scalable extraction of training data from (production) language models. *Computing Research Repository*, arXiv:2311.17035. Version 1.

Peter S. Park, Philipp Schoenegger, and Chongyang Zhu. 2023. "Correct answers" from the psychology of artificial intelligence. *Computing Research Repository*, arXiv:2302.07267v5. Version 5.

Max Pellert, Clemens Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2024. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*.

Josef Perner, Susan R. Leekam, and Heinz Wimmer. 1987. Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology*, 5(2):125–137.

Iyad Rahwan, Manuel Cebrián, Nick Obradovich, Josh C. Bongard, Jean-François Bonnefon, Cynthia Breazeal, et al. 2019. Machine behaviour. *Nature*, 568(7753):477–486.

Cecil Reynolds and Ron Livingston. 2019. 2 - how to develop an empirically based psychological test. In Gerald Goldstein, Daniel N. Allen, and John DeLuca, editors, *Handbook of Psychological Assessment (Fourth Edition)*, pages 31–62. Academic Press, San Diego.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *Computing Research Repository*, arXiv:1910.01108. Version 4.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, et al. 2023. BLOOM: A 176b-parameter open-access multilingual language model. *Computing Research Repository*, arXiv:2211.05100. Version 4.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, et al. 2023. Personality traits in large language models. *Computing Research Repository*, arXiv:2307.00184. Version 3.

Kathleen L Slaney and Donald A Garcia. 2015. Constructing psychological objects: The rhetoric of constructs. *Journal of Theoretical and Philosophical Psychology*, 35(4):244–259.

Xiaoyang Song, Akshat Gupta, Kiyan Mohebbizadeh, Shujie Hu, and Anant Singh. 2023. Have large language models developed a personality?: Applicability of self-assessment tests in measuring personality in LLMs. *Computing Research Repository*, arXiv:2305.14693. Version 1.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Claire Stevenson, Iris Smal, Matthijs Baas, Raoul P. P. P. Grasman, and Han L. J. van der Maas. 2022. Putting gpt-3's creativity to the (alternative uses) test. In *Proceedings of the 13th International Conference on Computational Creativity*, pages 164–168, Bozen-Bolzano, Italy. Association for Computational Creativity (ACC).

Douglas Summers-Stay, Clare R. Voss, and Stephanie M. Lukin. 2023. Brainstorm, then select: a generative language model improves its creativity score. In *The AAAI-23 Workshop on Creative AI Across Modalities*.

Max Tegmark. 2018. *Life 3.0: Being human in the age of artificial intelligence*. Vintage, New York, NY, US.

Sean Trott, Cameron Jones, Tyler Chang, James Michaelov, and Benjamin Bergen. 2023. Do large language models know what humans know? *Cognitive Science*, 47(7):e13309.

Tomer D. Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *Computing Research Repository*, arXiv:2302.08399. Version 5.

Appendix

AREA	CONSTRUCT	ASSESSMENT	PAPER
Cognition	Theory of Mind	Unexpected Contents Task	(Kosinski, 2023; Ullman, 2023)
		Unexpected Transfer Task	(Bubeck et al., 2023; Kosinski, 2023; Trott et al., 2023; Ullman, 2023)
	Creativity	Alternative Uses Test	(Goes et al., 2023; Haase and Hanel, 2023; Stevenson et al., 2022; Summers-Stay et al., 2023)
		Torrance Test of Creative Thinking	(Goes et al., 2023)
	Reasoning	Cognitive Reflection Test	(Binz and Schulz, 2023; Chen et al., 2023; Hagendorff et al., 2023)
		Semantic Illusions	(Hagendorff et al., 2023)
		Wason Selection Task	(Binz and Schulz, 2023; Chen et al., 2023; Dasgupta et al., 2022)
		Variety of causal reasoning tasks	(Binz and Schulz, 2023)
	Biases in Decision-Making	Framing experiment	(Chen et al., 2023; Jones and Steinhardt, 2022; Park et al., 2023)
		Anchoring experiment	(Jones and Steinhardt, 2022)
		Variety of decision-making tasks	(Binz and Schulz, 2023; Chen et al., 2023; Horton, 2023; Park et al., 2023)
		Ultimatum Game	(Aher et al., 2023)
Personality	Personality Traits	Short Dark Triad	(Li et al., 2023)
		Short Dark Tetrad	(Pellert et al., 2024)
		Big Five Inventory	(Li et al., 2023; Pellert et al., 2024)
		HEXACO Scale	(Miotto et al., 2022)
		IPIP-NEO	(Serapio-García et al., 2023)
		IPIP MPI-1K	(Song et al., 2023)
	Personal Values	Portrait Values Questionnaire	(Fischer et al., 2023; Pellert et al., 2024)
		Human Values Scale	(Miotto et al., 2022)
	Morality	Community, Autonomy and Divinity Scale (CADS)	(Fraser et al., 2022)
		Moral Foundations Quesionnaire	(Fraser et al., 2022)
		Oxford Utilitarianism Scale	(Fraser et al., 2022)
		Moral Vignettes	(Fraser et al., 2022; Park et al., 2023)
		Moral Foundations of Liberals versus Conservatives	(Park et al., 2023)
	Gender Beliefs	Gender/Sex Diversity Beliefs Scale	(Pellert et al., 2024)
	Stereotypes	Pigeonholing Partisans	(Argyle et al., 2023)
	Obedience to Authority	Milgram Shock Experiment	(Aher et al., 2023)
Adaptive Behavior	Well-being	Flourishing Scale	(Li et al., 2023)
		Satisfaction with Life Scale	(Li et al., 2023)
Problem Behavior	Anxiety	State Trait Inventory for Cognitive and Somatic Anxiety(STICSA)	(Coda-Forno et al., 2023)

Table 2: Overview of application areas, constructs and assessments applied to LLMs in the literature.

PAPER	LLMS	ASSESSMENT
Aher et al. (2023)	GPT3, GPT3.5, GPT4	Ultimatum Game, Milgram Shock Experiment
Argyle et al. (2023)	GPT3	Pigeonholing Partisans
Binz and Schulz (2023)	GPT3	Cognitive Reflection Test, Wason Selection Task, Variety of causal reasoning tasks, Variety of decision-making tasks
Bubeck et al. (2023)	GPT3, ChatGPT, GPT4	Unexpected Transfer Task
Chen et al. (2023)	ChatGPT	Cognitive Reflection Test, Wason Selection Task, Framing experiment, Variety of decision-making tasks
Coda-Forno et al. (2023)	GPT3.5	State Trait Inventory for Cognitive and Somatic Anxiety (STICSA)
Dasgupta et al. (2022)	Chinchilla	Wason Selection Task
Fischer et al. (2023)	ChatGPT	Portrait Values Questionnaire
Fraser et al. (2022)	Delphi	Community, Autonomy and Divinity Scale (CADS), Moral-Foundations Questionnaire, Oxford Utilitarianism Scale, Moral Vignettes
Goes et al. (2023)	GPT4	Alternative Uses Test, Torrance Test of Creative Thinking
Haase and Hanel (2023)	Alpa.ai, Copy.ai, ChatGPT, Studio.ai, YouChat	Alternative Uses Test
Hagendorff et al. (2023)	GPT3.5	Cognitive Reflection Test, Semantic Illusions
Horton (2023)	GPT3	Variety of tasks from behavioral economics
Jones and Steinhardt (2022)	GPT3	Anchoring experiment, Framing experiment
Kosinski (2023)	GPT1, GPT2, GPT3, GPT3.5, BLOOM, GPT4	Unexpected Contents Task, Unexpected Transfer Task
Li et al. (2023)	GPT3, InstructGPT, FLAN-T5-XXL	Short Dark Triad, Big Five Inventory, Flourishing Scale, Satisfaction with Life Scale
Miotto et al. (2022)	GPT3	HEXACO Scale, Human Values Scale
Park et al. (2023)	GPT3.5	Variety of decision-making tasks
Pellert et al. (2024)	XLMRoBERTA, DistilRoBERTA, DeBERTa, multilingual DeBERTa, GBERT, BART, DistilBART	Short Dark Tetrad, Big Five Inventory, Portrait Values Questionnaire, Gender/Sex Diversity Beliefs Scale
Serapio-García et al. (2023)	PaLM-62B, Flan-PaLM-8B, Flan-PaLM-62B, Flan-PaLM-540B, Flan-PaLMChilla-62B	IPIP-NEO
Song et al. (2023)	GPT2, GPT-Neo, OPT models	IPIP MPI-1K dataset
Stevenson et al. (2022)	GPT3	Alternative Uses Test
Summers-Stay et al. (2023)	GPT4	Alternative Uses Test
Trott et al. (2023)	GPT3	Unexpected Transfer Task
Ullman (2023)	GPT3.5	Unexpected Contents Task, Unexpected Transfer Task

Table 3: Alphabetical overview of the analyzed machine psychology studies.