https://doi.org/10.1038/s41562-024-02096-z

How to evaluate the cognitive abilities of LLMs

Anna A. Ivanova

Language models have become an essential part of the burgeoning field of artificial intelligence (AI) psychology. I discuss 14 methodological considerations that can be used to design more robust, generalizable studies that evaluate the cognitive abilities of language-based AI systems, as well as to accurately interpret the results of these studies.

Ever since the Turing test, the idea of having a dialogue with a machine to probe its cognitive abilities ('thought') has been inextricably associated with the field of AI. In addition to its intuitive simplicity, this idea naturally aligns with everyday practice in psychology: language-based assessments are the bread and butter of many psychologists' toolkits. If researchers want to know what is happening in the mind of a human, the easiest approach is to ask.

Today, advances in linguistic abilities of large language models (LLMs) — and AI systems that might incorporate LLMs as one of their components — make it possible to seamlessly test these models on language-based assessments originally designed for people. This is an unprecedented advance: to date, the only entities who could flexibly use human language were human. Now, however, we are faced with artificial systems that can process linguistic information, generate novel texts and respond to questions, which raises the question of how we might assess the cognitive capabilities of these systems.

Easy access to chat-based LLM interfaces (such as ChatGPT) makes it possible for anyone to run a 'cognitive test' on an AI system. This advance has led to an explosive growth of AI psychology, with papers that assess LLMs' working memory capacity, logical reasoning, planning abilities, social reasoning, creativity and even personality traits. Advances in vision–language models now also make it possible to test AI systems on cognitive assessments that incorporate pictures and videos.

Although running such assessments can be fairly straightforward, interpreting the results is not. In fact, AI psychology today is faced with a plethora of methodological questions, including what factors should be considered when assessing model performance; how we can adapt our stimuli to reduce the prevalence of 'hacks' (that is, heuristics that the model might use to achieve high task performance without using the cognitive skill being assessed); and whether, if a model passes a human test for a cognitive ability 'X', this means that it indeed possesses X.

To begin answering these questions, I provide a list of 14 'dos and do nots' to consider when designing, conducting and interpreting the results of an AI psychology study (see refs. 1,2 for additional considerations) (Table 1). I do not touch on the philosophical question of whether it is at all appropriate to ascribe mental capacities to a machine; my goal here is simply to clarify the methodological criteria that determine the inferences we can and cannot make on the basis of a model's responses to a questionnaire or a cognitive test.

Table 1 | The 'dos and do nots' of AI psychology

Do	Do not
Determine what the model might have learned about your test during training	Rely on well-known (or minimally changed) test items
Consider alternative strategies that a model might use to arrive at the correct answer	Overly trust crowdsourced or automatically generated items
Incorporate careful control conditions	Overly trust LLM item scoring
Evaluate models under conditions similar to humans	Assume that model responses reflect 'universal' human behaviour
Examine the effect of culture-specific aspects of the prompt on model behaviour	Assume that models solve the task in the same way as humans
Compare model and human performance	Jump to conclusions
Be explicit about the experimental settings when reporting the results	
Check whether a model generalizes beyond a single test	

Experimental design

When designing or adapting a cognitive test for evaluating an AI model, it is important to consider the following points.

Do determine what the model might have learned about your test during training. The two most important issues to consider are whether the model was directly trained on your task and whether the model 'saw' examples from your test during its training. In the worst case, these issues might occur together — that is, the model was trained on the task of interest using the same items as those in the current study.

Do consider alternative strategies that a model might use to arrive at the correct answer. Even if the model has not directly learned your task during training or finetuning, it might still use a different strategy from the strategy you, the experimenter, have presupposed. The most prominent shortcuts for LLMs include word associations based on their statistical co-occurrence (for example, 'race car' and 'fast') — although there might be many others, such as grammatical cues or more abstract structural patterns in text. Humans, too, routinely use task strategies that are not initially considered by the experimenters, but the kinds of alternative strategies available to humans and to AI systems are not necessarily the same. Considering possible shortcut paths to solutions can help to both design shortcut-proof items and identify edge cases in which shortcut-based models are likely to fail[3].

Do incorporate careful control conditions. Once you identify alternative strategies by which a model might solve the task, the next step is to design control conditions to help to distinguish these possible strategies: for instance, whether the model will choose answer 'a' over

BOX 1

The Winograd schema challenge

In 2012, Levesque and colleagues proposed a test for diverse kinds of basic world knowledge[6] (named after an initial example by T. Winograd). The test includes minimally different item pairs with pronouns whose referent can be inferred on the basis of general world knowledge:

(1) Question: the trophy doesn't fit into the brown suitcase because it's too small. What is too small? Answer: the suitcase.
(2) Question: the trophy doesn't fit into the brown suitcase because it's too large. What is too large? Answer: the trophy.

A set of such minimally differing sentence pairs that tackle various world knowledge phenomena — physical properties, biology, social situations and so on — was then compiled into the Winograd schema challenge (WSC). Although initially challenging, the WSC was largely solved by 2020 and LLMs achieved over 90% accuracy[7]. However, it turned out that this success did not reflect models' deep knowledge of the world, but rather the flaws in the test itself.

In the original paper that introduced the challenge[6], the authors cautioned against using examples that could be solved via simple heuristics. They cite two such heuristics: selectional restrictions and frequency effects. The selectional restrictions heuristics can be illustrated with the example: "The women stopped taking the pills because they were pregnant/carcinogenic". Here, only an animate entity can be pregnant and only an inanimate entity can be carcinogenic, which leads to an unambiguous association between the adjectives and the corresponding nouns without the need to tap into deeper world knowledge. The frequency heuristics applies to cases such as: "The race car zoomed by the school bus because it was going so fast/slow". A simple association between 'race car' and 'fast' will suffice to determine the referent. Finally, the authors note that the examples need to be, in their words, 'Google-proof' — that is, not present in online text corpora used to train the models.

Despite the field's awareness of these heuristics, constructing a heuristics-free dataset was hard. In the original WSC dataset, over 10% of the items turned out to have simple association-based solutions[8]. To address this issue, Sakaguchi et al.[7] constructed a large dataset called WinoGrande, in which the examples were first crowdsourced online and then automatically filtered to reduce cooccurrence-based associations as estimated through language model embeddings (rather than human intuition). This filtering substantially reduced model performance, which suggests that the association heuristic indeed inflates model accuracy.

Soon after, Elazar et al.[9] showed that adding even more-stringent quality controls leads to substantial decreases in model performance on the WSC. The authors introduced two baseline cases to account for the raw frequency of possible continuations: sentences with candidate referents excluded ("doesn't fit into because it's too large/small") and sentences with the first half excluded ("because it's too large/small"). The assumption is that for these reduced sentences, there should be no consistent preference for 'suitcase' versus 'trophy' and — if there is — this reflects an association bias. It turned out that LLMs performed above chance on the WSC even in those baseline conditions, which indicated previously undetected association biases.

The final issue raised by the WSC story is the quality versus quantity trade-off in test design. The initial WSC set includes fewer than 300 examples, all of which were handcrafted by scholars. Now that these examples are freely available online, performance on this set is no longer reflective of the models' general capabilities. The authors of WinoGrande used a different approach: to obtain thousands of novel items, they asked human workers to come up with many different examples and then used automatic filtering techniques. This approach is more scalable but suffers from numerous quality issues, for example, typos ('wit' instead of 'with'), grammatical errors ('more brighter color') and unwarranted assumptions ('good at math' means 'likely to be a professor'). Overall, the trade-off between result generalizability (which is harder for small datasets) and quality control (which is harder for large datasets) remains an important issue to consider in test design.

Kocijan et al.[10] formulate several lessons from the WSC saga, the most important of which is perhaps: "[w]e need to be careful not to rely on a perceived connection between tasks and methods". Just because we think that a certain task requires a cognitive ability 'X', does not mean that it actually does.

'b' simply because 'a' is more frequent, or whether the model will show the same preference for answer 'a' over 'b' even when the question is omitted.

Do not rely on well-known (or minimally changed) test items. When tested on a famous test item (such as 'The trophy did not fit the suitcase because it was too small') (Box 1), the model has probably seen it multiple times during training and therefore has a high chance of relying on answer memorization; thus, its performance may not generalize to new, non-famous test items. Even if the experimenter makes a minor change to the original item (such as replacing 'trophy' with 'prize'), a model might still be able to complete the test on the basis of simple association between the old and the new item.

Do not overly trust crowdsourced or automatically generated items. The ability of AI models to process large amounts of queries — far more than any human — creates the temptation to evaluate their performance on thousands of test items, obtained from online human workers, created by automatic item-generation scripts or even generated by AI models themselves. However, if the quality of these items is not rigorously evaluated, such tests may be meaningless. AI-generated items present their own set of challenges: a model might generate test items that resemble samples from the training data or, more broadly, generate patterns that have a high likelihood under the model and might therefore facilitate its performance. A systematic assessment of such biases in AI-generated materials remains to be done.

Do evaluate models under conditions similar to humans when comparing their performance. When conducting direct model-to-human comparisons, the default should be to recreate the same testing conditions for humans and models[4]. If a human needs instructions to perform well on a test, the model should also receive those instructions. If a human performs well on a task with no in-the-moment training, a model with human-level performance would also need to perform the task with no in-context training. If divergences during evaluation occur (for either theoretical or practical reasons), they need to be justified and highlighted when presenting the results.

Do examine the effect of culture-specific aspects of the prompt on model behaviour. The language in which the test is presented, the dialect, the vocabulary used and even the use of capitalization and punctuation might all affect AI model performance, as it leverages these cues to mimic specific users online.

Interpreting the results

When interpreting the results of a cognitive assessment, the following factors are important.

Do compare model and human performance. We often assume humans will perform well on a specific test even when we do not have direct evidence for it. If these exact test items and question or prompt wordings have not been tested on humans before, they should be.

Do be explicit about the experimental settings when reporting the results. Computational work should, ideally, have perfect replicability. For that to be achievable, all the materials and code needed to run the experiment should be made available online. For models offered by commercial third parties, there is no guarantee that they will continue to offer access to the model, so some backup arrangements may need to be made. Finally, a third party may roll out a model update that would change experimental results; in such cases, the date of the experiment needs to be reported as well.

Do not overly trust LLM item scoring. The flip side of testing a model on thousands of items is the need to evaluate responses to all these items. In some cases, determining correct answers is straightforward (for example, a–d for a multiple-choice question); however, scoring free response items is hard work. Many studies are opting for LLM-based item evaluation; however, such approaches might be circular because the items that are hard for the model being evaluated might also be hard for the model doing the evaluation. At a minimum, the model being tested and the model doing the evaluation should belong to different classes; and even then, evaluator model performance needs to be rigorously checked by comparing it with human evaluations across a range of item difficulties.

Do not assume that model responses reflect 'universal' human behaviour. Even when directly comparing AI models and humans, it is important to be specific about which human population serves as a reference. Human psychology suffers from over-reliance on WEIRD (Western, educated, industrialized, rich and democratic) participant pools; AI psychology suffers from similar issues[5] because WEIRD individuals disproportionately contribute to the training data. Thus, calls to replace human data in psychology studies with AI-generated responses need to account for the strong demographic and cultural biases that these models bring. Similarly, model performance in English (the most represented language on the Internet) might not be reflective of model performance in other languages.

Do not assume that models solve the task in the same way as humans. Even if both humans and models do well on a particular test, it does not mean that they solve it in the same way. Some approaches to evaluating similarities and differences in the mechanisms that underlie human and model task performance include comparing their error patterns, generalization to new tasks, and the internal (neural) representations required to accomplish the task.

Do check whether a model generalizes beyond a single test. Even when we control for possible shortcuts, a model might still find a loophole that allows it to perform well on a particular test. However, the more diverse tests we include (and the more we control for the shortcuts in each), the harder it will be to attribute model performance to serendipitous factors. If a model does well on 20 logical reasoning tasks that vary in their content and format, it is more plausible that it possesses logical reasoning than if it performs well on one such task.

Do not jump to conclusions. The discourse around AI models has become very polarized, with both extreme enthusiasm based on a few isolated examples and extreme scepticism based on isolated failures. What the field needs is careful evaluation of specific model capabilities with a clear acknowledgement of advances and limitations. Moreover, results reported for one model may not generalize to other models, so conclusions from a single study should be qualified accordingly.

Testing open versus closed models

Some of the 'dos' discussed above — checking the training data for contamination with test items, verifying that the model has not been finetuned on the exact task being tested, and creating conditions that will allow the study to be replicated in the future on the exact same model — are essentially impossible in the case of closed models.

Thus, there is an argument to stop running AI psychology assessments on closed models altogether, given that – from a scientific perspective – the results are potentially non-reproducible and difficult to interpret.

And yet, it is unlikely that researchers will stop running cognitive tests on closed AI models. Although scientifically questionable, evaluating the behaviour of closed, industry-standard models has important practical implications that include understanding which real-life tasks they can (and cannot) be reliably used for; what AI models are in principle capable of achieving (given that closed models often exhibit state-of-the-art performance); and what safety risks they may present. Thus, going forward, cognitive evaluations of AI systems might be used in two separate settings: basic scientific inquiry into the cognitive capacities of AI systems (which should prioritize open models) and applied studies that tackle questions related to model performance, safety and user impact (which can be applied both to open and to closed models, with the caveat that results from closed models might not replicate or generalize).

As with any rapidly growing research area, the scientific practices and norms in AI psychology are constantly being defined and redefined, often through trial and error. I hope that this discussion of the 'dos and do nots' of AI psychology will help to distil some of the lessons learned and improve the robustness and validity of future work that aims to probe the cognitive capacities of AI systems.

Anna A. Ivanova
Georgia Institute of Technology, Atlanta, GA, USA.
e-mail: a.ivanova@gatech.edu

Published online: 15 January 2025

References

1. Frank, M. C. *Nat. Rev. Psychol.* **2**, 451–452 (2023).
2. Mitchell, M. *Science* **381**, eadj5957 (2023).
3. McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D. & Griffiths, T. L. *Proc. Natl Acad. Sci. USA* **121**, e2322420121 (2024).
4. Lampinen, A. K. Preprint at https://doi.org/10.48550/arXiv.2210.15303 (2022).
5. Atari, M., Xue, M. J., Park, P. S., Blasi, D. & Henrich, J. Prepint at PsyArXiv https://doi.org/10.31234/osf.io/5b26t (2023).
6. Levesque, H., Davis, E. & Morgenstern, L. The Winograd schema challenge. In *Proc. Thirteenth International Conf. on the Principles of Knowledge Representation and Reasoning* (eds Brewka, G. et al.) 552–561 (AAAI Press, 2012).
7. Sakaguchi, K., Bras, R. L., Bhagavatula, C. & Choi, Y. *Commun. ACM* **64**, 99–106 (2021).
8. Trichelair, P., Emami, A., Trischler, A., Suleman, K. & Cheung, J. C. K. How reasonable are common-sense reasoning tasks: a case-study on the Winograd schema challenge and SWAG. In *Proc. 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th International Joint Conf. on Natural Language Processing (EMNLP-IJCNLP)* (eds. Inui, K. et al.) 3382–3387 (ACL, 2019).
9. Elazar, Y., Zhang, H., Goldberg, Y. & Roth, D. Back to square one: zrtifact detection, training and commonsense disentanglement in the Winograd schema. In *Proc. 2021 Conf, on Empirical Methods in Natural Language Processing* (eds Moens, M.-F. et al.) 10486–10500 (ACL, 2021).
10. Kocijan, V., Davis, E., Lukasiewicz, T., Marcus, G. & Morgenstern, L. *Artif. Intell.* **325**, 103971 (2023).

Acknowledgements

I gratefully acknowledge funding support from the University System of Georgia. I thank A. Sathe, B. Lipkin, C. Kauf, G. Tuckute and K. Mahowald for their constructive comments.

Competing interests

The author declares no competing interests.

Additional information

Peer review information *Nature Human Behaviour* thanks Michael Frank, Adele Goldberg and Yoshua Bengio for their contribution to the peer review of this work.