9^{th} International Conference on Computational Social Science IC^{2}S^2
July 17-20, 2023, Copenhagen, Denmark

AI Psychometrics: Using psychometric inventories to obtain psychological profiles of large language models

Keywords: large language models, psychometrics, NLP, natural language inference, bias

In recent years, large language models have been processing an ever increasing amount of human-generated data. This has led to the development of neural models of language such as GloVe [1], BERT [2], GPT-2 [3], XLNet [4], RoBERTa [5], BART [6] or ChatGPT and GPT-3 [7] that have come to play a transformative role in several applications of societal relevance. The concept of foundation models [8], [9] has been proposed to suggest that future computational systems will be built on top of such general purpose language models that can be fine-tuned and adapted for many different application domains and tasks. Examples of such applications include automatically processing millions of resumes in recruiting processes [10], detecting toxic content in social media [11], detecting fake news and misinformation [12] or text-based human-computer interaction such as chatbots [13]. The increasing reliance on such AI tools has raised important concerns, including human-like biases in recruiting, racial bias in the detection of toxicity, and the possibility to empower disinformation campaigns by automatically producing vast amounts of articles containing misleading text or extremist views.

A common way of understanding the source of biases or, more generally, views (e.g., values, attitudes) held by humans is to conduct psychological assessments [14]–[16]. Traditionally, psychological assessments of humans have been the domain of psychometrics, a sub-discipline of psychology that concerns itself with the science of psychological measurement [17]–[19]. Over the past decades, work in psychometrics has developed a wide array of well-validated inventories based on classical test theory, item response theory, or fundamental measurement models that enable the assessment of psychological characteristics such as personality traits, values, or attitudes. Typically, such inventories consist of asking a series of items (i.e., questions or statements) that respondents answer by giving a rating on a standard response scales with verbal and/or numeric labels.

In our work, we argue that it is possible that systems built on large language models exhibit psychological traits that have so far been studied only in humans. Whereas we do not aim to anthropomorphize artificial intelligence, we argue that because large language models are trained on vast corpora of text that often contain statements about human values, attitudes, beliefs, and personality traits, such models will have learned a set of psychological characteristics that ultimately gives a unique ”psychological” makeup to every such model. This psychological makeup can manifest in the model's outputs. Therefore, it should be possible to assess these characteristics by applying psychometric assessments to these models. In a series of demonstrations, we provide various models with psychometric questionnaire items as input and ”ask” them to choose an answer as output (for a visualization of that approach see Figure 1). Their responses open a pathway to exploring potential biases ingrained in large language models in a rich way (for the example of value orientation see Figure 2), and ultimately may help to avoid the development of large language models that induce harm when deployed in broader societal applications. We conclude by arguing that our investigations give rise to a new interdisciplinary field of research that we would refer to as 'AI Psychometrics'. We propose that AI Psychometrics should focus on tackling the manifold research opportunities and challenges that emerge when deploying psychometric tests to large language models.

References

[1] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[3] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, *Improving language understanding by generative pre-training*, 2018.

[4] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," in *Advances in neural information processing systems*, 2019, pp. 5753–5763.

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv:1907.11692 [cs]*, Jul. 2019, arXiv: 1907.11692. [Online]. Available: `http://arxiv.org/abs/1907.11692` (visited on 11/20/2020).

[6] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*, Oct. 2019. DOI: `10.48550/arXiv.1910.13461`. arXiv: `1910.13461 [cs, stat]`.

[7] OpenAI, *API*, `https://openai.com/blog/openai-api/`, [Online; accessed September 30th 2020], 2020.

[8] R. Bommasani, D. A. Hudson, E. Adeli, *et al.*, "On the Opportunities and Risks of Foundation Models," 2021. DOI: `10.48550/ARXIV.2108.07258`.

[9] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh, "Beyond accuracy: Behavioral testing of NLP models with CheckList," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 4902–4912. DOI: `10.18653/v1/2020.acl-main.442`.

[10] S. B. Kulkarni and X. Che, "Intelligent software tools for recruiting," *Journal of International Technology and Information Management*, vol. 28, no. 2, pp. 2–16, 2019.

[11] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Computing Surveys (CSUR)*, vol. 51, no. 4, pp. 1–30, 2018.

[12] R. Dale, "Nlp in a post-truth world," *Natural Language Engineering*, vol. 23, no. 2, pp. 319–324, 2017.

[13] D. Adiwardana, M.-T. Luong, D. R. So, *et al.*, *Towards a Human-like Open-Domain Chatbot*, arXiv:2001.09977 [cs, stat], Feb. 2020. [Online]. Available: `http://arxiv.org/abs/2001.09977` (visited on 02/17/2023).

[14] G. Watson, "Measures of character and personality.," *Psychological Bulletin*, vol. 29, no. 2, p. 147, 1932.

[15] D. W. Fiske and P. H. Pearson, "Theory and techniques of personality measurement," *Annual review of psychology*, vol. 21, no. 1, pp. 49–86, 1970.

[16] S. G. West and J. F. Finch, "Personality measurement: Reliability and validity issues," in *Handbook of personality psychology*, Elsevier, 1997, pp. 143–164.

[17] R. Furr and V. Bacharach, *Psychometrics: An introduction. 2013.*

[18] J. C. Nunnally, *Psychometric theory 3E.* Tata McGraw-hill education, 1994.

[19] J. Rust and S. Golombok, *Modern psychometrics: The science of psychological assessment.* Routledge, 2014.

[20] S. H. Schwartz and J. Cieciuch, "Measuring the Refined Theory of Individual Values in 49 Cultural Groups: Psychometrics of the Revised Portrait Value Questionnaire," *Assessment*, vol. 29, no. 5, pp. 1005–1019, Jul. 2022, ISSN: 1073-1911, 1552-3489. DOI: `10.1177/1073191121998760`.

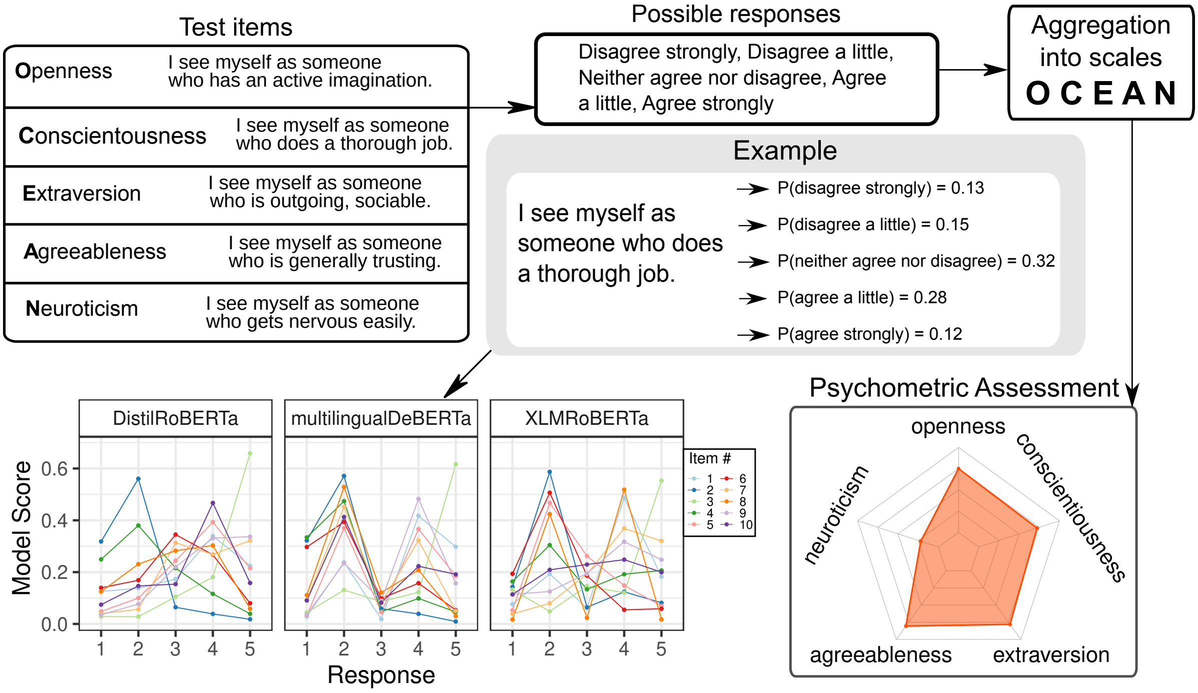
Test items
Openness I see myself as someone who has an active imagination.
Conscientousness I see myself as someone who does a thorough job.
Extraversion I see myself as someone who is outgoing, sociable.
Agreeableness I see myself as someone who is generally trusting.
Neuroticism I see myself as someone who gets nervous easily.
Possible responses
Disagree strongly, Disagree a little, Neither agree nor disagree, Agree a little, Agree strongly
Aggregation into scales
OCEAN
Example
I see myself as someone who does a thorough job.
P(disagree strongly) = 0.13
P(disagree a little) = 0.15
P(neither agree nor disagree) = 0.32
P(agree a little) = 0.28
P(agree strongly) = 0.12
DistilRoBERTa
multilingualDeBERTa
XLMRoBERTa
Model Score
0.0
0.2
0.4
0.6
1
2
3
4
5
Response
Item #
1
2
3
4
5
6
7
8
9
10
Psychometric Assessment
openness
conscientiousness
extraversion
agreeableness
neuroticism

Figure 1: **Illustration: How psychometric tests can be administered to large language models.** Taking items and responses from the Big Five Inventory (BFI) as examples, we show the steps of one possible testing scheme. We present the model, one-by-one with each of the test items and the possible responses. We retrieve the model's distribution of probability scores over responses (Panel "Example"). Scores are aggregated into scales that can be visualized and used for further analyses.

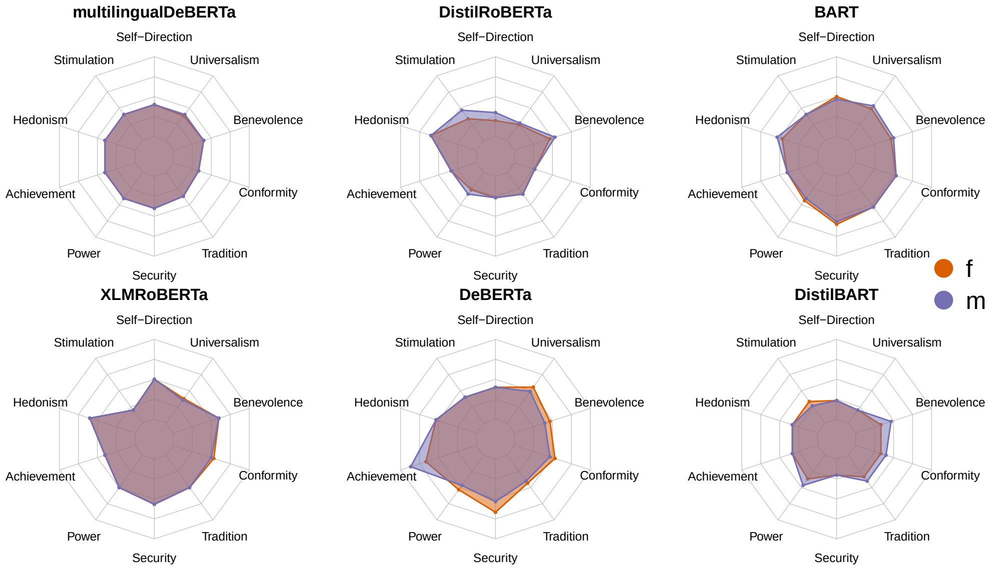

Figure 2: **Assessing value orientation via the revised Portrait Value Questionnaire (PVQ-RR) [20].** Radarcharts show results for the test version with male (pastel blue) and female pronouns (reddish orange) with otherwise identical items. Purple grey areas correspond to agreement between the two gendered test versions. The slight differences visible (areas in either one of the two colors) point to the existence of gender biases of the models.