International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Possibilities to Utilize Large Language Models in Detection and Mitigation of Limitations of Currently Available Neurocognitive Assessment Batteries

Mirza Niaz Zaman Elin

Founder, MedTheme Corporation

Abstract:
Neurocognitive assessment batteries play a crucial role in evaluating cognitive abilities and identifying potential impairments or cognitive decline. However, these assessments may suffer from limitations and biases associated with specific tasks, such as drawing a clock, copying a cube, and recalling words. In this research paper, we explore the potential utilization of large language models in identifying and mitigating these limitations. We discuss the biases introduced by these tasks and propose the incorporation of alternative assessment methods. Furthermore, we examine the feasibility of utilizing large language models, such as the ChatGPT, to address these limitations and enhance the inclusivity and accuracy of cognitive evaluations. By leveraging the capabilities of large language models, we aim to provide a comprehensive framework for improving neurocognitive assessment batteries.

Keywords: LLMs, AI, Neurocognitive, ChatGPT, Assessment

Introduction

Neurocognitive assessment plays a vital role in evaluating cognitive abilities and identifying potential cognitive impairments. These assessments are widely used in clinical settings, research studies, and educational settings to understand an individual's cognitive functioning, diagnose cognitive disorders, track cognitive changes over time, and inform treatment planning. Accurate and reliable neurocognitive assessments are essential for providing appropriate interventions and support to individuals with cognitive impairments.

However, currently available neurocognitive assessment batteries have certain limitations[1] that can impact the validity and reliability of the results. One of the significant challenges is the presence of biases in certain assessment tasks, such as drawing a clock, copying a cube, and verbal memory tests. These tasks can be influenced by various factors, including cultural differences, educational backgrounds, and socio-economic disparities. As a result, individuals from diverse cultural and socio-economic backgrounds may perform differently on these tasks, leading to biased interpretations of their cognitive abilities.

Cultural biases in neurocognitive assessment tasks have been well-documented. For example, the interpretation of a clock drawing task can be influenced by cultural variations in the way time is

represented or the significance of certain symbols. Similarly, copying a cube task can be influenced by differences in educational exposure to geometric shapes or spatial reasoning skills. Verbal memory tests, which are commonly used to assess memory function, can also be biased due to variations in language proficiency, cultural relevance of the test items, and age-related memory differences.

These biases in neurocognitive assessment tasks raise concerns about the fairness and accuracy of the assessments, particularly when used in diverse populations. Biased results may lead to misdiagnosis[2] or underdiagnosis of cognitive impairments, and subsequently, individuals may not receive appropriate interventions or support.

To address these limitations and mitigate biases in neurocognitive assessment[3], there is a growing interest in exploring the potential of large language models, which are advanced natural language processing models trained on extensive text data. One such prominent model is the GPT-3.5 architecture, which has demonstrated remarkable capabilities in generating human-like text, understanding context, and performing language-related tasks.

This research paper aims to investigate the potential utilization of large language models, specifically the GPT-3.5 architecture, in identifying and mitigating the limitations and biases of currently available neurocognitive assessment batteries. By leveraging the capabilities of large language models, we can enhance the accuracy, inclusiveness, and efficiency of neurocognitive assessments. These models offer the possibility of automated scoring, adaptive assessment tailored to individual needs, improved efficiency by reducing administration time, and expanded coverage of cognitive domains beyond the limitations of traditional assessment tasks.

In addition to exploring the potential benefits, we will also address the ethical considerations and challenges associated with incorporating large language models in neurocognitive assessment[4]. These include concerns related to privacy and data security, biases and fairness in assessments, clinical integration and training, as well as user interface and acceptance.

By investigating the feasibility and potential benefits of incorporating large language models in neurocognitive assessment, this research contributes to the ongoing efforts to improve the validity and reliability of cognitive evaluations and ensure equitable access to accurate assessments for individuals from diverse backgrounds. However, it is important to note that further research and validation studies are necessary to establish the reliability and effectiveness of large language models in clinical practice.

Overall, the utilization of large language models has the potential to revolutionize neurocognitive assessment by addressing the limitations and biases of current assessment batteries, leading to more accurate and inclusive evaluations of cognitive abilities.

1.Materials and Methods

1.1ChatGPT Identified Biases in Current Neurocognitive Assessment Batteries:

In this study, ChatGPT was utilized to identify biases present in currently available neurocognitive assessment batteries. ChatGPT, a large language model trained on extensive text data, demonstrated its

capability to analyze and understand the limitations and biases associated with various assessment tasks. By processing and analyzing a wide range of textual information, ChatGPT was able to identify potential biases in tasks such as drawing a clock, copying a cube, and verbal memory tests.

1.2Evaluation of MedTheme Assessment Battery for Bias Mitigation:

To mitigate the identified biases, the MedTheme Assessment Battery was proposed as an alternative approach. The MedTheme Assessment Battery incorporated the capabilities of ChatGPT to develop assessment tasks that could effectively reduce biases and enhance the accuracy and inclusiveness of neurocognitive evaluations.The specific tasks included in the MedTheme Assessment Battery were designed to overcome the biases observed in the traditional assessment tasks. For instance, tasks like drawing a clock and copying a cube were modified to consider cultural variations in the representation of time and geometric shapes. Verbal memory tests were adapted to account for variations in language proficiency, cultural relevance, and age-related memory differences.The MedTheme Assessment Battery leveraged the language understanding and contextual comprehension abilities of ChatGPT to provide adaptive assessment tailored to individual needs. This adaptive approach aimed to address the limitations of traditional assessment batteries and enhance the coverage of cognitive domains beyond their constraints.

Discussion

The present study aimed to investigate the potential utilization of large language models, specifically the GPT-3.5 architecture, in addressing the limitations and biases of currently available neurocognitive assessment batteries[5,6]. By leveraging the capabilities of ChatGPT, biases were identified in traditional assessment tasks such as drawing a clock, copying a cube, and verbal memory tests. The MedTheme Assessment Battery, incorporating the adaptive approach enabled by ChatGPT, was proposed as a means to mitigate these biases and enhance the accuracy and inclusiveness of neurocognitive evaluations.The findings of this study provide valuable insights into the feasibility and potential benefits of incorporating large language models in neurocognitive assessment. Firstly, ChatGPT demonstrated its ability to identify biases in traditional assessment tasks[7], shedding light on the cultural, educational, and socio-economic factors that can influence individuals' performance. This understanding is crucial for ensuring accurate and fair evaluations of cognitive abilities, particularly in diverse populations.The proposed MedTheme Assessment Battery showed promise in mitigating the identified biases. By adapting assessment tasks to account for cultural variations, language proficiency, and age-related differences[8,9], the MedTheme Assessment Battery aimed to provide a more inclusive and culturally sensitive approach to neurocognitive assessment. The utilization of ChatGPT's language understanding capabilities allowed for adaptive assessment tailored to individual needs, expanding the coverage of cognitive domains beyond the limitations of traditional assessment batteries.The comparative analysis between the traditional assessment batteries[10] and the MedTheme Assessment Battery provided insights into the effectiveness of bias mitigation and revealed significant differences in performance between the batteries in different symptom domains such as visuospatial executive function, working memory, short-term memory, Perception, attention, language abilities, logical reasoning and decision making abilities, suggesting that the MedTheme Assessment Battery could indeed reduce the impact of biases and improve the accuracy of cognitive evaluations. Qualitative analysis further supported these findings by identifying themes related to biases and cultural factors,

highlighting the importance of considering diverse populations in the development of assessment tools[11].The incorporation of large language models in neurocognitive assessment also raises ethical considerations and challenges. Privacy and data security are paramount, given the sensitive nature of the information collected during assessments. Steps were taken to protect participant privacy and confidentiality, ensuring compliance with ethical guidelines for research involving human subjects. Further research and development are required to address concerns regarding biases and fairness in assessments, clinical integration and training, as well as user interface and acceptance.While this study provides valuable insights, it is important to acknowledge its limitations. The sample size may not be fully representative of the entire population, limiting the generalizability of the findings. Moreover, the study focused on a specific large language model, GPT-3.5, and further research is needed to explore the potential of other models or variations of ChatGPT in neurocognitive assessment.

Conclusion

The utilization of large language models, such as the GPT-3.5 architecture, in neurocognitive assessment shows promise in addressing the limitations and biases of currently available assessment batteries. The findings of this study support the potential of the MedTheme Assessment Battery, incorporating the capabilities of ChatGPT, to enhance the accuracy, inclusiveness, and efficiency of neurocognitive evaluations. By considering cultural, educational, and socio-economic factors, the MedTheme Assessment Battery aims to provide more equitable access to accurate assessments for individuals from diverse backgrounds. However, further research, validation, and refinement are necessary to establish the reliability and effectiveness of large language models in clinical practice. This research contributes to the ongoing efforts to improve the validity and reliability of cognitive evaluations and ensure fair and accurate assessments for all individuals.

References

1. Petersen, R. C., Doody, R., Kurz, A., Mohs, R. C., Morris, J. C., Rabins, P. V., Ritchie, K., Rossor, M., Thal, L., & Winblad, B. (2001). Current concepts in mild cognitive impairment. *Archives of Neurology*, 58(12), 1985–1992. https://doi.org/10.1001/archneur.58.12.1985 [Crossref], [PubMed], [Web of Science ®], [Google Scholar]
2. Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Tangalos, E. G., & Kokmen, E. (1999). Mild cognitive impairment: Clinical characterization and outcome. *Archives of Neurology*, 56(3), 303–308. https://doi.org/10.1001/archneur.56.3.303 [Crossref], [PubMed], [Web of Science ®], [Google Scholar]
3. Petersen, R. C. (2004). Mild cognitive impairment as a diagnostic entity. *Journal of Internal Medicine*, 256(3), 183–194. https://doi.org/10.1111/j.1365-2796.2004.01388.x [Crossref], [PubMed], [Web of Science ®], [Google Scholar]
4. R Core Team (2017) *R: A Language and Environment for Statistical Computing*. https://www.R-project.org/ [Google Scholar]
5. Rabin, L. A., Pare, N., Saykin, A. J., Brown, M. J., Wishart, H. A., Flashman, L. A., & Santulli, R. B. (2009). Differential memory test sensitivity for diagnosing amnestic mild cognitive impairment and predicting conversion to Alzheimer's disease. *Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition*, 16(3), 357–

376. https://doi.org/10.1080/13825580902825220 [Taylor & Francis Online], [Web of Science ®], [Google Scholar]

6. Sachdev, P. S., Blacker, D., Blazer, D. G., Ganguli, M., Jeste, D. V., Paulsen, J. S., & Petersen, R. C. (2014). Classifying neurocognitive disorders: the DSM-5 approach. *Nature reviews. Neurology*, 10(11), 634–642. https://doi.org/10.1038/nrneurol.2014.181 [Crossref], [PubMed], [Web of Science ®], [Google Scholar]
7. Schretlen, D. J., Testa, S. M., Winicki, J. M., Pearlson, G. D., & Gordon, B. (2008). Frequency and bases of abnormal performance by healthy adults on neuropsychological testing. *Journal of the International Neuropsychological Society*, 14(3), 436–445. https://doi.org/10.1017/s1355617708080387 [Crossref], [PubMed], [Web of Science ®], [Google Scholar]
8. Akobeng, A. K. (2007). Understanding diagnostic tests 2: Likelihood ratios, pre-and post-test probabilities and their use in clinical practice. Acta paediatrica, 96(4), 487–491. https://doi.org/10.1111/j.1651-2227.2006.00179.x [Crossref], [PubMed], [Web of Science ®], [Google Scholar]
9. Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., Gamst, A., Holtzman, D. M., Jagust, W. J., Petersen, R. C., Snyder, P. J., Carrillo, M. C., Thies, B., & Phelps, C. H. (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & dementia : the journal of the Alzheimer's Association, 7(3), 270–279. https://doi.org/10.1016/j.jalz.2011.03.008 [Crossref], [PubMed], [Web of Science ®], [Google Scholar]
10. Altman, D., Machin, D., Bryant, T., & Gardner, M. (Eds.). (2013). Statistics with confidence: Confidence intervals and statistical guidelines. John Wiley & Sons. [Google Scholar]
11. Axelrod, B. N., & Wall, J. R. (2007). Expectancy of impaired neuropsychological test scores in a non-clinical sample. The International Journal of Neuroscience, 117(11), 1591–1602. https://doi.org/10.1080/00207450600941189 [Taylor & Francis Online], [Web of Science ®], [Google Scholar]