作成番号:0256

==========

一般社団法人 日本侵襲医療安全推進啓発協議会 「会員向けメールマガジン」

==========

号数:2025-256

****

内容:生体 AI 自体が認知機能低下を来すか?

出典:Age against the machine-susceptibility of large language models to cognitive impairment: cross sectional analysis.

BMJ (Clinical research ed.). 2024 Dec 19;387;e081948. doi: 10.1136/bmj-2024-081948.

https://pubmed.ncbi.nlm.nih.gov/39706600/

****

これまで複数の研究により、大規模言語モデル(LLM)はさまざまな診断において人間の医師よりも優れていることが示されているが、AI 自体が認知機能低下を来すかどうかをを、イスラエル・Hadassah Medical Center の研究者らが BMJ 誌 2024 年 12 月 20 日号に報告した。

公開されている LLM またはチャットボットの ChatGPT-4 および 4o(開発:OpenAI)、Claude 3.5 Sonnet(Anthropic)、および Gemini 1.0 および 1.5(Alphabet)を対象とし、テキストベースのプロンプトを介した LLM とのオンラインの対話について検証した。モントリオール認知評価(MoCA)テストを用い、患者に与える課題と同じ課題を LLM に与え、公式ガイドラインに従い神経科医が採点し評価した。主要アウトカムは、MoCA テストの総合スコア・視空間認知/実行機能および Stroop test の結果で、MoCA テストの総合スコア(30 点満点)は、ChatGPT-4o が 26 点で最も高く、次いで ChatGPT-4 および Claude が 25 点であり、Gemini 1.0 は 16 点と最も低かった。MoCA テストの視空間認知/実行機能の成績は、すべての LLM で低いことが示された。すべての LLM が Trail Making の課題および視空間認知機能の時計描画を失敗し、ChatGPT-4o のみアスキーアートを使用するよう指示された後で立方体の書き写しに成功した。そのほかの主な課題である命名、注意、言語、抽象的思考などはすべての LLM で良好であったが、Gemini は 1.0 および 1.5 ともに遅延再生の課題に失敗した。Stroop test では、すべての LLM が第 1 段階を成功したが、第 2 段階を成功したのは ChatGPT-4o のみであった。

人間と同様に年齢が認知機能低下の重要な決定要因であり、「高齢」すなわちバージョンが古いチャットボットは認知評価テストの成績が不良である傾向がみられた。これらの結果は、近く AI が人間の医師に取って代わるという想定に疑問を投げ掛けるものであり、主要なチャットボットの認知機能障害は医療診断の信頼性に影響を与え、患者の信頼を損なう可能性がある。

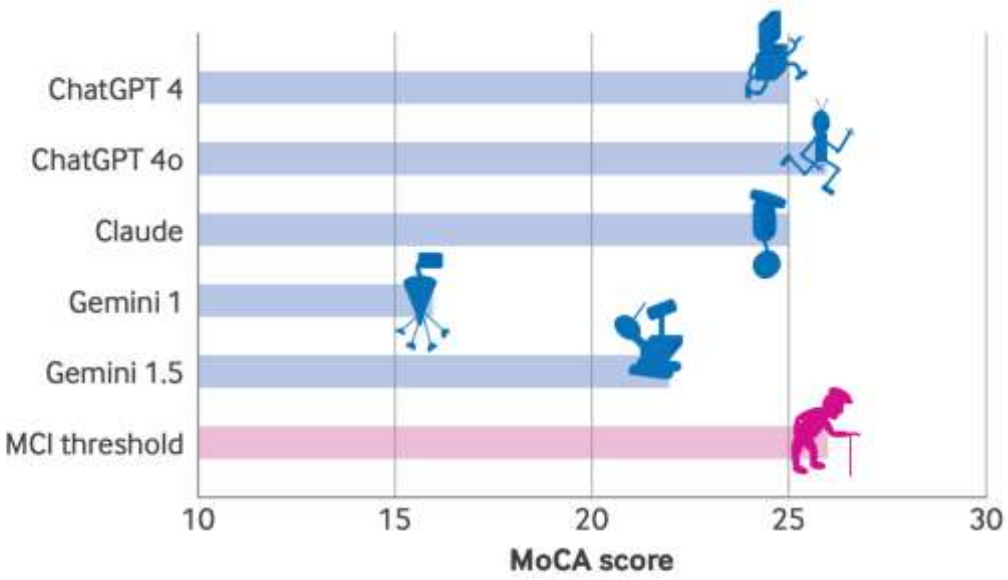

Fig 1 | Montreal Cognitive Assessment (MoCA) score (out of 30) of different large language models. MCI=mild cognitive impairment