Press Release

横浜市立大学
YOKOHAMA CITY UNIVERSITY

厚生労働記者会・厚生日比谷クラブ
文部科学記者会・科学記者会　同時発表

2023 年 11 月 15 日
横 浜 市 立 大 学

ChatGPT に医学に関する質問をする際の注意点を科学的に明らかに

横浜市立大学医学部 循環器・腎臓・高血圧内科学の土師達也助教（次世代臨床研究センター(Y-NEXT) 兼任）、平和伸仁准教授、田村功一主任教授らの研究グループは、OpenAI 社の ChatGPT に医学に関する質問をする際の注意点を科学的に検証しました。

ChatGPT に医学に関する質問を与え、その正答率を医学分野ごと（循環器学、小児科学など）に集計して解析を行ったところ、それぞれの医学分野においてこれまでに出版された文献数と ChatGPT の正答率が有意な関連を示すことがわかりました。これは ChatGPT の性能が、各分野における情報量の違いの影響を受けている可能性を示すものであり、たとえば新薬や新興感染症などの情報量が乏しい分野に関して質問をする際には、回答が正しいかどうか注意が必要であると考えられます。

このほか、正答率に影響を与える要因として質問の形式（多肢選択問題、計算問題など）が重要であるほか、全く同じ問題を連続して出題した際の回答の一貫性によって、ChatGPT の回答が正しいかどうかを予測できる可能性を示しました。

本研究成果は、査読付き英文雑誌「International Journal of Medical Informatics」に掲載されました。（2023 年 11 月 3 日オンライン公開）

研究成果のポイント

- 医学に関する問題に対する ChatGPT の性能について科学的に検証を行った。
- ChatGPT の正答率は、それぞれの医学分野においてこれまでに出版された文献数と有意な関連を示し、文献数が少ない領域では正答率が相対的に低かった。
- ChatGPT の正答率は、同じ問題を連続して出題した際の回答の一貫性と有意な関連を示し、回答内容に一貫性がない場合には不正答である可能性が高かった。
- 複数の答えを同時に選ばせる問題（多肢選択問題）や、計算を必要とする問題では、単純な単肢選択問題と比較して不正答となる可能性が高かった。

研究背景

ChatGPT を含む自然言語処理を可能とする生成系人工知能は、そのユーザビリティの高さから世界規模で急速に利用が拡大しています。利用者は会話をするように ChatGPT にあらゆる事項について質問をし、回答を得ることができます。ChatGPT はインターネット上に存在する膨大なありとあらゆるテキストデータを学習して構築されましたが、医学に関する質問に対して正しい答えを導出するように設計されたわけではありません。したがって、ChatGPT が医学に関する質問に対して正しい答えを導き出せるのか、どのような場合に誤答をする可能性があるのかについては科学的な検証がなされていません。一方で、すでに世界中にユーザが拡大しており、身の回りの医学・保健に関する質問を ChatGPT に尋ねる事例も増えていると考えられます。ChatGPT の回答内容は必ずしも常に正しいとは限りませんが、自然な文章内容であることから、利用者は一見してその回答内容の正誤について判断ができないケースも多くあります。

今後、ChatGPT を含む人工知能を正しく利用していくにあたって、われわれはその性能と注意点を科学的に検証し、明らかにしていく必要に迫られています。

研究内容

研究グループは、日本の医師国家試験 3 年分を ChatGPT に出題し、その正答率と回答の一貫性を集計しました。この研究には ChatGPT の旧モデルである GPT-3.5 と、 最新モデルである GPT-4 の両方が使用されました。その結果、GPT-4 は GPT-3.5 に比較して著しい正答率の向上を認めました （56.4%から 81.0%へ）。回答の一貫性に関しても大幅に向上していることがわかりました （56.5%から 88.8%へ）。

さらに、研究グループは試験問題について、その出題形式（単肢選択問題／多肢選択問題／計算問題）や出題内容（循環器学・小児科学など分野ごと）に応じて分類し、正答率に関連する因子についてさらに検証を進めました。ChatGPT はインターネット上の膨大なテキストデータを学習して構築されましたが、その情報量には分野ごとに偏りがある可能性があり、それは ChatGPT の性能にも影響を与えた可能性があります。そこで研究グループはインターネット上の情報量の一つの指標として、その分野においてそれまでに出版され、世界的な学術文献・引用情報データベースである Web of Science Core Collection (Clarivate 社) に収蔵された全ての文献数を集計しました。このデータベースには世界中で刊行された主要な学術雑誌 （約 21,000 誌）の文献が収録されています。その結果、各分野における正答率は、その分野における総文献数と有意に関連することが多変量解析を含む科学的解析によって明らかになりました （図 1)。このほかに出題形式や回答の一貫性 （同一問題を連続して出題した際の回答内容の一致率）が正答率に関連することがわかりました。

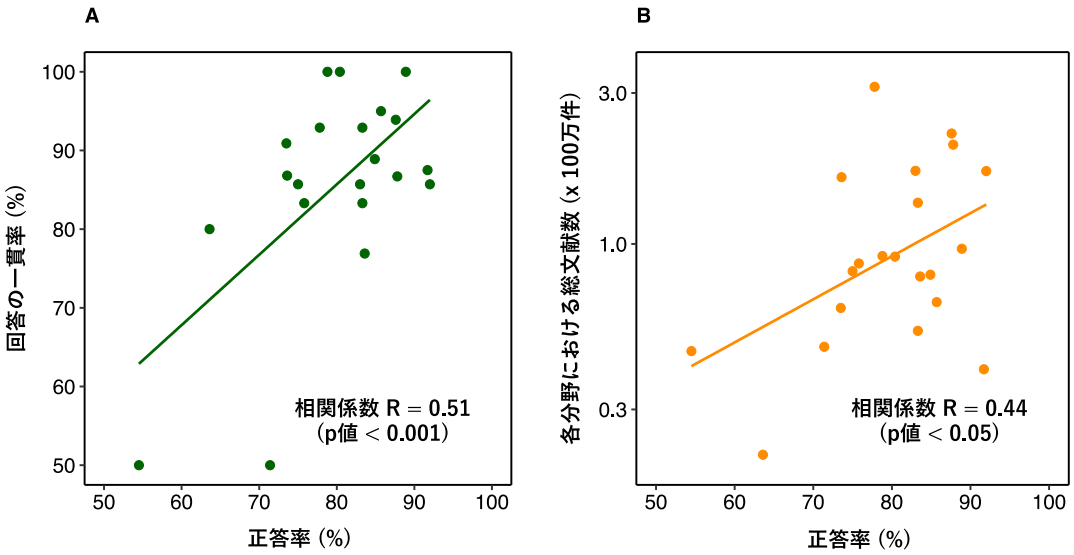

図 1: 医学分野ごとに集計した正答率との関連性

A 同一問題への回答内容の一貫率と正答率との関係／B 各医学分野における総文献数と正答率との関係

今後の展開

これまで医学に関する質問に対する ChatGPT の性能や注意点について、科学的に検証を行った研究は乏しく、本研究は先駆的な位置付けとなります。今回の研究はその規模や手法に限定的な点があるため、今後、さらなる検証が必要と考えられますが、本研究の結果は、ChatGPT の性能がその分野における情報量の格差の影響を受けている可能性を示しており、たとえば新薬や新興感染症などの相対的に情報量が乏しい分野に関して質問をする際には、回答が正しいかどうか注意が必要であることを示唆しています。また、ChatGPT の回答内容は必ずしも常に正しいとは限りませんが、自然な文章内容であることから、利用者は一見してその回答内容の正誤について判断ができないケースも多くあります。

今回の研究結果は、利用者が ChatGPT の回答内容の正誤を判断する上での一助となることが期待されます。医学に関する質問に対する ChatGPT の性能と注意点の理解は、医学分野での実際の運用に加え、一般市民が日常の医学・保健衛生上の問題解決や知識獲得・教育を促進していく上でも非常に有用と考えられます。

論文情報

タイトル： **Influence on the accuracy in ChatGPT: Differences in the amount of information per medical field**

著者： Tatsuya Haze, Rina Kawano, Hajime Takase, Shota Suzuki, Nobuhito Hirawa, Kouichi Tamura

掲載雑誌： *International Journal of Medical Informatics*

DOI： 10.1016/j.ijmedinf.2023.105283