配信先：文部科学記者会、科学記者会、千葉県政記者クラブ、PR Times

ニュースリリース

千葉大学
CHIBA UNIVERSITY

令和６年７月９日
国立大学法人千葉大学

大規模言語モデルによる文献スクリーニングの効率化

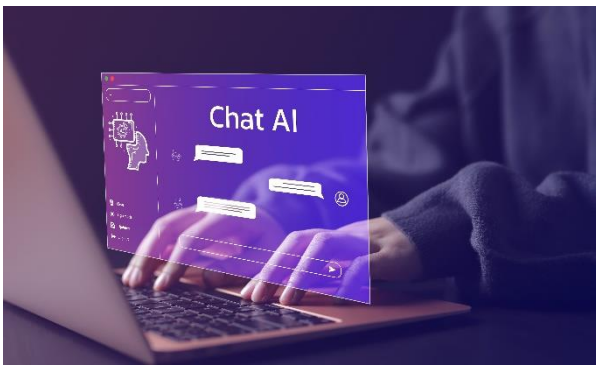
―診療ガイドライン作成における AI の活用―

千葉大学医学部附属病院 救急集中治療医学の大網毅彦講師、同大学院医学研究院の中田孝明教授、国立シンガポール大学 Duke-NUS Medical School の岡田 遥平研究員らの研究チームは、ChatGPT などの大規模言語モデル（LLM: Large Language Model）[注 1)] を診療ガイドライン作成の一部に用いることで、ガイドライン作成に必要な文献を膨大な医学情報の中から高い精度で見つけ出すことができることがわかりました。同時に、医学文献検索にかかる膨大な作業時間を従来の方法の 10 分の 1 以下まで短縮できることも明らかになりました。

医師や看護師などの医療従事者が中心となって作成する診療ガイドラインには多くの人手や時間が必要です。現在、医師の働き方改革[注 2)] が進められており、医療従事者の労働負担を減らすことが重要な課題となっています。AI を活用した効率的な文献スクリーニング方法は、持続可能な働き方を実現するための一つの解決策として期待されます。

日本版敗血症診療ガイドライン[注 3)] 2024 作成委員会の取り組みの一環として行われた本研究成果は、総合医学雑誌 JAMA Network Open に 2024 年 7 月 8 日（現地時間）に掲載されました。

■研究の背景



診療ガイドラインは、ある疾患に対する検査や治療を決めるための道標として医療従事者や患者さんが参考にする文書です。このガイドラインを作成するために必要なシステマティックレビューという作業は、ある医学領域に関連する文献を抽出し、文献の情報を同定、選択や評価を行う作業で、多くの労力や時間を要します。一方、人工知能（AI）の一種である ChatGPT などの LLM は学習した大量のデータをもとに、人間が指示した命令や質問に答えることができます。この LLM が、システマティックレビュー作業の中でも特に多くの労力を要する文献の抽出作業を代わりに行うことができれば、人間が行うべき作業量を大幅に削減することができます。しかしこれまで、LLM を用いた文献スクリーニング作業の精度や作業負担軽減の程度は検討されていませんでした。

本研究では、日本版敗血症診療ガイドラインの作成において、LLM を用いた文献スクリーニングの精度と効率性を評価しました。

■研究内容と結果

本研究は、診療ガイドラインの中から 5 つの臨床疑問（CQ）に関する文献スクリーニングデータを使用して、LLM（OpenAI より 2023 年 11 月 7 日に公開された GPT-4 Turbo）がそれぞれの CQ に関連するキーワードをもとに抽出された数多くの文献の中から、CQ に含まれる患者/集団/問題、介入、比較、および研究デザインに合致する文献を正確に選び出すことができるかどうかを検証しました。LLM の文献スクリーニングの正確性を評価するために、ガイドラインメンバーが実際に行った文献スクリーニングの結果をゴールドスタンダード[注 4)] として、LLM を使用した文献スクリーニングの結果を評価しました。具体的な正確性の指標として、感度[注 5)] と特異度[注 6)] を計算しました。また、従来のスクリーニング方法と LLM を用いたスクリーニング法の作業時間を比較しました。

① 大規模言語モデルを用いた文献スクリーニングの正確性

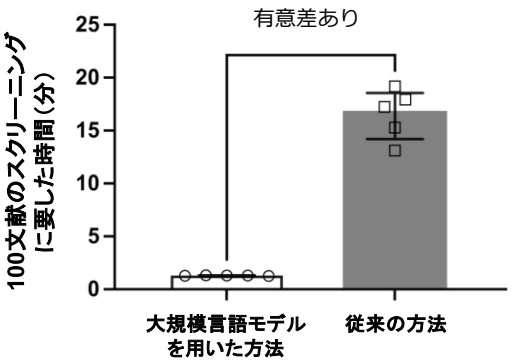
・従来の文献スクリーニング方法では、CQ1 で 5,634 件中 8 件、CQ2 で 3,418 件中 4 件、CQ3 で 1,038 件中 4 件、CQ4 で 4,326 件中 17 件、CQ5 で 2,253 件中 8 件がガイドラインを作成するための最終的な文献として選定されました。5 つの CQ における主要解析では、LLM を用いたスクリーニングの感度は 0.75（95% 信頼区間[CI]、0.43–0.92）、特異度は 0.99（95% CI, 0.99–0.99）でした。LLM の特徴として、人間が入力する命令文（コマンドプロンプト）によって LLM の作業内容が変化することが報告されています。本研究において、LLM における作業の質が向上するようにコマンドプロンプトを修正したところ、感度は 0.91（95%

CI, 0.77–0.97）に上昇し、特異度はほとんど低下しませんでした（0.98 ; 95% CI, 0.96–0.99）。

②大規模言語モデルを用いた文献スクリーニングの作業時間の短縮

・LLM を用いた文献スクリーニングは、2～4 人のガイドラインメンバーが人力で文献スクリーニングを行う従来の方法では 17.2 分かかっていた 100 件の文献スクリーニング時間を、約 1.3 分に短縮しました（平均差 −15.25 分、95% CI, −17.70～−12.79）（図 2）。

図 2 文献スクリーニングに要した時間の比較



■今後の展開

今回の結果から、LLM を用いた文献スクリーニングはある程度の正確性を有していること（許容できる感度と非常に高い特異度）がわかりました。また、文献スクリーニングにかかる時間を大幅に短縮しました。この新しい文献スクリーニングの方法は、システマティックレビューの効率を向上させ、作業負担を軽減する可能性があります。

現在様々な LLM の開発が行われており、その性能や機能は日進月歩です。今後発表される改良版の LLM を用いることによって、文献スクリーニングの精度がさらに高まることが予想されます。また、今回の研究の中で検討したコマンドプロンプトは学問として発達途上にあるため、今後の知見によって LLM を用いた作業内容の正確性が大いに高まる可能性もあります。このように LLM は今後さらに文献スクリーニングの精度や作業負担を改善する可能性があり、注目されていく分野であると考えられます。今回検討したのは敗血症の分野のみの文献検索でしたが、その他の医学分野においても LLM を用いた文献スクリーニングの応用が期待されます。医療従事者の作業負担を減らしながら、より良い診療ガイドラインを作成するために、今後も AI を活用した作業の効率化につながる取り組みを続けていきます。

■用語解説

注１）大規模言語モデル（LLM）：膨大な文章データを学習して、人間のように文章を理解したり作成したりする AI 技術。ChatGPT はその一例で、質問に答えたり、文章を生成したりすることができる。

注 2）医師の働き方改革：長時間労働の是正や、有給休暇の取得促進、多様な働き方の実現を目指した医師の労働環境の改善を目的とした一連の取り組み。

注 3）日本版敗血症診療ガイドライン：敗血症は、感染症が原因で全身に強い炎症が起こり、命に関わる状態になる病気。本ガイドラインは、世界の専門家が集まって作成した、敗血症診療ガイドライン ‘Surviving Sepsis Campaign Guidelines 2021’ を参考にして日本の現状に沿って作成したもの。

注 4）ゴールドスタンダード：特定の疾患や条件の診断、治療、評価において最も信頼性が高いと認められている方法または基準。医学や科学の分野で広く用いられ、最も確かな証拠に基づく選択肢として推奨される。

注 5）感度：ある検査が病気を持っている人を正しく見つけ出す能力を指す。感度が高いほど、病気を見逃さないという意味であり、例えば、感度が 100%であれば、病気を持っている人を全員見つけ出すことができる。

注６）特異度：ある検査が病気を持っていない人を正しく除外する能力を指す。特異度が高いほど、健康な人を誤って病気だと判断する可能性が低いことを意味する。例えば特異度が 100%であれば、健康な人を全員正しく健康だと判断することができる。

■論文情報

題名: Performance of a large language model in screening citations
著者名: Takehiko Oami, Yohei Okada, Taka-aki Nakada
掲載誌: JAMA Network Open

<研究に関するお問い合わせ>
千葉大学医学部附属病院救急集中治療医学 講師 大網 毅彦（おおあみ たけひこ）
TEL: 043-226-2372　メール: ccya0318@chiba-u.jp
URL: https://www.ho.chiba-u.ac.jp/hosp/section/kyukyu/index.html
<広報に関するお問い合わせ>
千葉大学広報室
TEL: 043-290-2018　メール: koho-press@chiba-u.jp