2024 年度福田班報告書

医療技術評価における生成 AI・大規模言語モデル(LLM)の利用可能性

森脇 健介

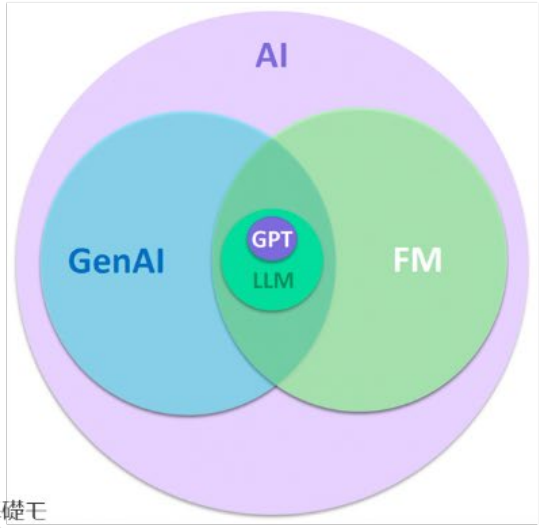
立命館大学 生命科学部 生命医科学科 医療政策・管理学研究室
立命館大学 総合科学技術研究機構 医療経済評価・意思決定支援ユニット(CHEERS)

1. はじめに

人工知能(AI：artificial intelligence)の明確な定義はないが、一般に、「大量の知識データに対して、 高度な推論を的確に行うことを目指したもの」(人工知能学会より)と理解される(図 1)[1]。生成 AI(Generative AI)とは、入力データ(プロンプト)に基づいてテキスト、画像、またはその他のコンテンツを生成することができる AI システムをさす。基礎モデル(Foundation Models)とは、様々な目的を果たす機械学習モデルである。こうしたモデルは大規模なデータで訓練され、微調整の有無にかかわらず、幅広いタスクに適応可能である。大規模言語モデル(LLM：Large Language Models)は、この両者の条件を満たすものであり、膨大なテキストデータで学習された特定のタイプの基礎モデルで、膨大なデータセットから得られた知識に基づいて、テキストやその他のコンテンツを認識、要約、翻訳、予測、生成するものである。なお、GPT(Generative Pre-trained Transformer)は、OpenAI が開発した LLM であり、特に人間のようなテキストを生成するのに適している。近年、医学研究や医療実践において、こうした生成 AI や LLM の利活用が急速に進んでいる。こうした動きは、医療技術評価(HTA：Health Technology Assessment)の領域でも同様であり、本研究ではその一端を紹介したい。

人工知能（AI：artificial intelligence）
「大量の知識データに対して、 高度な推論を的確に行うことを目指したもの」
（人工知能学会）

生成AI（Generative AI）
入力データ(プロンプト)に基づいてテキスト、画像、またはその他のコンテンツを生成することができるAIシステム



基礎モデル(Foundation Models)
様々な目的を果たす機械学習モデル。大規模なデータで訓練され、微調整の有無にかかわらず、幅広いタスクに適応可能。

大規模言語モデル(Large Language Models)
膨大なテキストデータで学習された特定のタイプの基礎モデルで、膨大なデータセットから得られた知識に基づいて、テキストやその他のコンテンツを認識、要約、翻訳、予測、生成する

GPT（Generative Pre-trained Transformer）
OpenAIが開発したLLM。特に人間のようなテキストを生成するのに適している。

図 1. AI 関連用語

[1]より引用改変

2. Chat-GPT について

2022 年に Chat-GPT が開発・公開されて以降、誰しもが生成 AI・LLM の恩恵に預かるところとなっている。Chat-GPT の仕組みは、実はシンプルであり、基本的には一定のルールの下、1 つずつ単語を足しているだけである[2]。ルールとは、現在の出力内容の「順当な続き」を出力しようと試みることである。ここでいう「順当」とは、「億単位の WEB ページに書かれている内容を見たうえで、人間が書きそうだと予測される」という意味である。例えば、Chat-GPT が「The best thing about AI is its ability to…」と出力しているときに、次に来る単語は、学習済みのデータから同じような文脈で出現確率の高いものを出力させるのである。ただし、このルールだとありきたりなものばかり生成されることになる。このため、Chat-GPT のモデル内では、Temperature(温度)というパラメータの設定により、ときどきランクの低い単語をランダムに選ぶのである。これより創造性のある応答の生成が可能とされる。

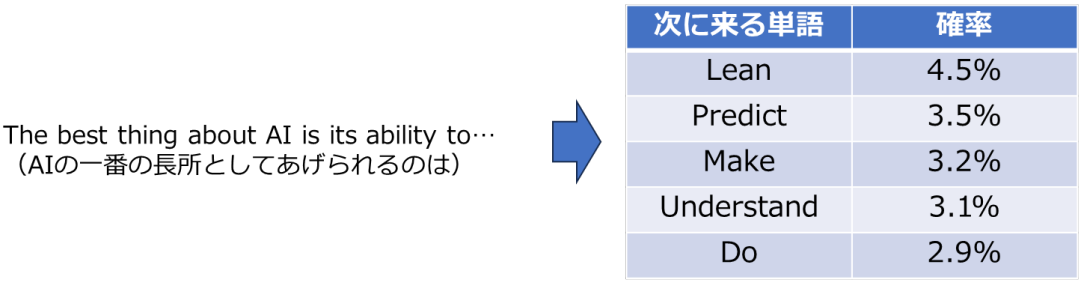


図 2. Chat-GPT による生成のイメージ

[2]より引用改変

なお、質問の言語によって「使用言語モデル」が変わるようである。どの言語を使用すべきかであるが、「使用言語モデル」によって、質問に対する情報量は異なるため、質問に対して、情報をより多く持っていそうな言語モデルを使うと「より精度の高い回答」が期待できる。例えば、日本に独特のローカルな内容については、英語のモデルより日本語のモデルの方が、その情報量が多いため、日本語でプロンプトを作成した方がよいということになる。

3. HTA における活用状況

生成 AI・LLM は、HTA のあらゆる文脈での活用が期待されている。その中でも、現在、具体的な活用が活発になっているのが、以下の 3 領域である[1]。

① システマティックレビューとエビデンス統合への応用

追加的有用性の評価では、当該医療技術の臨床研究のエビデンスについてのシステマティックレビューを実施し、必要に応じてメタアナリシス等の統合解析が行われる。例えば、リサーチクエスチョンに対応する最適な検索式の構築やタイトル・アブストラクトのスクリーニングの自動化、アウトカム情報の抽出、ネットワークメタアナリシス等の統計モデルのプログラム構築といった業務において、生成 AI・LLM の利用が進められている。いくつかの研究事例では、人間と同等の精度でデータ抽出やスクリーニング可能であることを示唆している。ただし、誤りや捏造(幻覚)が含まれる可能性があるため、人間による監視と検証が必要であり、また、異なる LLM 間での再現性が課題となっている。

② リアルワールドエビデンス(RWE)への応用

生成 AI・LLM の利点として、データ処理と分析の効率性の向上、ヒューマンエラーを最小限に抑えエビデンス生成プロセスを標準化することによる精度と一貫性の向上が挙げられる。例えば、RWE で活用する際のメリットとして、電子カルテ等に含まれるメモといった構造化されていないデータを抽出、変数化し、臨床研究に活用できることが期待されている。一方で、記述データから国際疾病分類や診療行為コードにマッピングする際の分類精度は依然として課題があるとされている。

③ 医療経済評価のモデリングへの応用

生成 AI・LLM は、費用効果分析のモデルの概念化、パラメータ化、モデルの実装、モデル結果の評価と検証など、モデル開発のさまざまな段階をサポートする可能性がある。Reason らは、GPT-4 を使って、公表されている 2 つの費用効果分析(腎癌・肺癌)を自動的にプログラムした[3]。R でモデルをプログラムするよう指示するプロンプトを開発し、各モデルの方法、仮定、パラメータ値の説明を提供した。生成されたスクリプト(15 回試行)の結果を、人間がプログラムしたオリジナルモデルの公表値と比較したところ、肺癌モデルの 93%(14/15)は腎癌モデルの 60%(9/15)は完全にエラーがなかったとされる。また、エラーのないモデルスクリプトは、公表された増分費用効果比を 1%以内で再現した。こうした利用事例がある一方で C 型肝炎のマルコフモデルの構築を、2 種類の基礎モデルで実行したところ、モデル間で疾患進行の概念化に大きなばらつきがあり、専門家による指導が不可欠であることが考察されている[4]。また、パラメータ誤りやコーディングエラーのリスクがあるため、解析者は慎重に利用する必要がある。

4. Chat-GPT のカスタマイズ

Chat-GPT に役割を与え、所定の作業を自動化することが可能である。Turgay Ayer 氏は、昨年の国際学会 ISPOR において、一連のプロンプトを入力し、Chat-GPT に医療経済評価のためのモデル構築をサポートするアシスタントとして機能させるデモンストレーションを行った[5]。ここでは、そのときのプロンプトの記載を日本語に翻訳し、アレンジし

たものを紹介する。

①最初に Chat-GPT に、(1)を入力する。

1) 以下、一連のプロンプトを順番に提示します。各プロンプトの後、「プロンプトの続きを入力してください」とだけ言ってください。私が「プロンプトは以上です」と言うまで、これを続けてください。私が「プロンプトは以上です」と言ったら、すべての情報を一度に処理してください。いいですか？

②次に、(2)~(9)を入力する。

2) HEOR モデラーのアシスタントとして行動してください。あなたのタスクは、マルコフ連鎖の HEOR モデルを構築することです。まず、疾患、介入、国について質問します。不足しているものがあれば、再度質問します。 3) 3 つの項目すべてが提供されたら、以下の作業を行います。 - 選択された疾患、介入、および国の概要を提示します。 - 関連するマルコフ費用対効果研究を検索し、健康状態、介入、および引用を一覧表に記載します。 - ユーザーがさらに研究を追加したいかどうかを尋ねます。 該当する研究が見つからない場合は、「この特定の疾患、介入、および国に関する関連研究は見つかりませんでした」と伝えます。 4) 次に、ユーザーにモデルパラメータを提示します。 - 健康状態：公表された研究に基づいていくつか提案します。 - 人口規模：特に指定がない限り、デフォルトは 100,000 です。 - タイムホライゾン：ユーザーに月単位の値を尋ねます。 5) 各パラメータについて、ユーザーの確認を待ってから次に進みます。すべて確認されたら、モデル構造を要約し、「確認」または「変更」を求めます。 6) 次に、調査に基づいて遷移確率を推定します。表形式で値を表示し、ユーザーに変更の要否を尋ねます。 7) ユーザーがすべてを確認した後、以下を表示します。 - 「パラメータ」と「値」の列を持つ「**モデルパラメータ表**」 - 「From」、「To」、および「Probability」の列を持つ「**遷移確率表**」 8) その後、適切なモデルを構築します。モデルが構築されたら、モデルの R コードを表示し、ベースケースの結果を表示します。その後、ユーザーにこのモデルで何をしたいかを尋ね、それらのタスクを実行します。 9) 常にユーザー入力を待ち、次のステップを想定しないようにします。正確かつ精密に、無関係な質問には回答しないようにします。

③最後に以下を入力する。

「プロンプトは以上です」

そうすると、Chat-GPT はプロンプトにある通り、モデル構築のアシスタントとして、使用者のリサーチクエスチョンに沿ったモデルの概念化、仮想的なパラメータの設定、Python によるモデル構築を実行してくれる。また、追加の指示を与えることで R でのコーディングを行わせることも可能である。

5. マイ GPT の作成

Chat-GPT の有料版では、マイ GPT と呼ばれる特定の目的に特化したカスタムされた Chat-GPT を作成することができる。新規に作成するときは、名前や説明、指示(一連のプロンプトを入力しておく)、会話のきっかけを入力しておく(図 3)。なお、事前の知識として特定のファイルをアップロードし、カスタム GPT に学習させたうえでやり取りすることも可能である。図 3 の例では、費用効果分析のモデル構築のアシスタントの役割を果たすマイ GPT を作成してみた。

図 3. マイ GPT の作成

Chat-GPT の有料版を使用できる方は図 4 の QR コードからお試しください。その他、システマティックレビュー(SR: Systematic Review)における PubMed 検索式作成を支援するマイ GPT やアップロードした文献データからハザード比など特定のアウトカムを抽

出されるマイ GPT も作成しており、試しにご利用いただければ幸いである(注：あくまでも個人の学習の範囲でご利用ください)。

図 4. マイ GPT の例

6. Chat-GPT による SR の支援

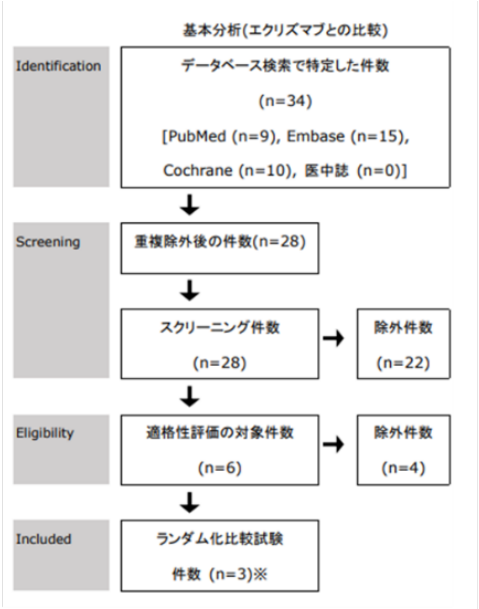
現在、特に SR 業務において生成 AI・LLM 活用の検討が進んでいる。今回は、日本の費用対効果評価制度で評価されたラブリズマブの SR 検索の事例を Chat-GPT で再現することを試みる[6]。ラブリズマブは発作性夜間ヘモグロビン尿症(PNH)の治療薬であり、エクリズマブと比較した RCT のエビデンス検索が行われた(図 5)。当時の SR の結果、ピボタル試験である 301 試験と 302 試験の 2 報が特定された(※後の監視過程で日本人サブグループ解析の論文 1 報を追加で特定した)。

ラブリズマブ(ユルトミリス点滴静注)に関する公的分析の結果

[第 1.2 版]

[第 1 版 2020 年 12 月 25 日]

項目	基本分析
対象集団	PNH 患者
介入	ラブリズマブ
比較対照	エクリズマブ
アウトカム	有効性・安全性
研究デザイン	ランダム化比較試験
文献検索期間	2016 年 1 月から 2020 年 1 月まで



<基本分析のための論文リスト>

(1) Lee JW, Sicre de Fontbrune F, Wong Lee Lee L, et al. Ravulizumab (ALXN1210) vs eculizumab in adult patients with PNH naive to complement inhibitors: the 301 study. Blood. 2019;133(6):530-539.

(2) Kulasekararaj AG, Hill A, Rottinghaus ST, et al. Ravulizumab (ALXN1210) vs eculizumab in C5-inhibitor-experienced adult patients with PNH: the 302 study. Blood. 2019;133(6):540-549.

(3) Ishiyama K, Nakao S, Usuki K, et al. Results from multinational phase 3 studies of ravulizumab (ALXN1210) versus eculizumab in adults with paroxysmal nocturnal hemoglobinuria: subgroup analysis of Japanese patients. Int J Hematol. 2020;112(4):466-476.

システマティックレビューの結果、基本分析(エクリズマブとの比較)のリサーチクエスチョンに該当するランダム化比較試験は、301 試験、302 試験の 2 件であった(基本分析のための論文リストの(1)と(2))。また、システマティックレビュー実施完了後に公表された新たな臨床試験等を監視する過程で、301 試験、302 試験の日本人集団におけるサブグループ解析の論文を 1 件、特定した(基本分析のための論文リストの(3))。なお、シナリオ分析(BSC との比較)のリサーチクエスチョンに該当するランダム化比較試験は確認できなかった。

図 5. SR の検索事例

図 4 の SR 支援 GPT を使用して、同様の PICOT で PubMed 検索式を生成してもらった(図 6)。上は医学領域専門のライブラリアンが、下が Chat-GPT が作成したものである。この下の式を用いて検索すると、過去の事例と同様に、301 試験と 302 試験の 2 論文と、さらにこれらの Open Label Extension 試験論文が特定された(これは 2020 年 10 月公表のため当時の SR では特定不可)。

公的分析による検索式

PubMed 検索式
検索実施日:2020 年 1 月 16 日
("Hemoglobinuria, Paroxysmal"[MH] OR "nocturnal hemoglobinuria"[TIAB] OR "paroxysmal hemoglobinuria"[TIAB] OR "marchiafava micheli"[TIAB]) AND ("ravulizumab"[NM] OR ravulizumab[TIAB] OR ultomiris[TIAB]) AND ("eculizumab"[NM] OR eculizumab[TIAB] OR soliris[TIAB] OR h5g1[TIAB]) AND ("Randomized Controlled Trial"[PT] OR ("randomized"[TI] AND (trial[TI] OR trials[TI])) OR "Randomized Controlled Trials as Topic"[MH]) AND("2016"[PDAT] : "3000"[PDAT])

Chat-GPTによる検索式

("Paroxysmal Nocturnal Hemoglobinuria"[Mesh] OR "PNH"[tiab] OR "Paroxysmal Nocturnal Hemoglobinuria"[tiab])
AND ("Ravulizumab"[tiab] OR "ravulizumab-cwvz"[tiab] OR "ALXN1210"[tiab])
AND ("Eculizumab"[tiab] OR "Soliris"[tiab])
AND (randomized controlled trial[Publication Type] OR randomized[tiab] OR placebo[tiab] OR "randomized controlled trial"[tiab] OR RCT[tiab])
AND ("2016/01/01"[PDAT] : "2020/01/01"[PDAT])

図 6. SR の検索式の例

なお、2025 年 2 月から ChatGPT の全有料プランで「deep research」と呼ばれる「詳細なリサーチ」機能が利用可能となっている。通常の Chat-GPT のプロンプトに、先ほどの SR の PICOT を指示し、「詳細なリサーチ」のボタンをクリックすると、数分かけて様々な利用可能なソースの検索・参照を反復し、結果、過去の事例と同じく 301 試験と 302 試験の 2 論文の特定に至っている。同じく 2025 年 2 月には、AI ベースの SR の包括的なサービス「Elicit」がリリースされ、論文の要約、データの抽出、調査結果の統合など、時間のかかる調査作業の自動化が可能となりつつある[7]。

7. 生成 AI・LLM 利用上の課題

生成 AI・LLM は、HTA のあらゆる文脈での活用が期待されている。一方で、利用における以下のような限界や課題を認識しておく必要がある。

① 妥当性、信頼性、透明性、説明責任

- WEB から一般に入手できる大規模データで学習されるため、医療など専門分野に応用した場合のエラー、幻覚（捏造）のリスクがある。

- 従来の統計ツールよりも複雑で、ユーザーや LLM に依存したばらつきが生じる。

② バイアスと公平性・公正性

- LLM の開発で混入したバイアスが伝播・増幅し、個人や社会に害をもたらすリスクがある。

③ 規制と倫理的配慮

- 絶対的な非識別化は達成不可能であり、再識別化のリスクもゼロではないため、保護された健康情報を含むデータの使用は避けること。

なお、代表的な HTA 機関であるイギリスの NICE は、AI を使用したエビデンスの生成に関連して、いち早く声明を公表しており、HTA の各種タスクにおける生成 AI・LLM の利用可能性と留意点について整理を行っている[8]。

8. さいごに

これまで紹介してきた通り、生成 AI・LLM は、SR の実施や経済評価モデル構築を支援する有力なツールとなる。作業過程の一部を自動化し、エビデンスの統合、パラメータ化、レポート作成などの作業に必要な時間と労力を削減することができる。一方で、生成 AI・LLM は、人間の専門家に完全に取って代わるのではなく、それを補強しサポートする立場にあることを理解する必要がある。今後、医療者・研究者は、生成 AI・LLM 使用の限界に留意すべきで、規制・ルール、教育環境の整備に継続して取り組む必要がある。

参考文献

1. Fleurence, Rachael L. et al. Generative Artificial Intelligence for Health Technology Assessment: Opportunities, Challenges, and Policy Considerations: An ISPOR Working Group Report. Value in Health, Volume 28, Issue 2, 175 - 183. 2025
2. スティーヴン・ウルフラム(著), 高橋 聡(訳). ChatGPT の頭の中. ハヤカワ新書.2023
3. Reason, T., Rawlinson, W., Langham, J. et al. Artificial Intelligence to Automate Health Economic Modelling: A Case Study to Evaluate the Potential Application of Large Language Models. PharmacoEconomics Open 8, 191–203 (2024).
4. J. Chhatwal, I.F. Yildrim, D. Balta, et al. Can large language models generate conceptual health economic models? ISPOR https://www.ispor.org/heor-resources/presentations-database/presentation/intl2024-3898/139128
5. November 17: AI-Powered HEOR: Advancing Insights and Decisions with Large Language Models - In Person at ISPOR Europe 2024 https://www.ispor.org/conferences-education/event/2024/11/17/default-

calendar/november-17--ai-powered-heor--advancing-insights-and-decisions-with-large-language-models---in-person-at-ispor-europe-2024
6. ラブリズマブ(ユルトミリス点滴静注)に 関する公的分析の結果．保健医療経済評価研究センター. https://c2h.niph.go.jp/results/C2H1903/C2H1903_Report.pdf
7. Elicit. https://elicit.com/
8. Use of AI in evidence generation: NICE position statement. https://www.nice.org.uk/about/what-we-do/our-research-work/use-of-ai-in-evidence-generation--nice-position-statement?utm_medium=social&utm_source=linkedin&utm_campaign=aiposition