別紙 3

厚生労働科学研究費補助金
政策科学総合研究事業（臨床研究等 ICT 基盤構築・人工知能実装研究事業）
総括研究報告書

研究課題名　大規模言語モデル（LLM：Large Language Model）を活用した医薬品等の有効性・安全性評価のためのアウトカム抽出の方法論の確立に向けた研究（24AC1004）

研究代表者　武藤 学　国立大学法人 京都大学大学院医学研究科　教授

研究要旨

本研究の初年度は、まず現在の法制度で医療向けの LLM を臨床データを使って研究開発し、実用化するための問題点について検討した。さらに LLM の基本的な適用範囲と能力の評価も行った。この段階では、経過記録等の非構造化情報（ダミー）を作成し、初期の LLM モデルを構築してトレーニングする。このプロセスには、テキストマイニング技術と最新の LLM のチューニング技術を統合することが含まれる。初期データ分析を通じて、モデルのパフォーマンスを評価し、抽出されたデータの有用性と精度を京大病院・腫瘍内科学講座を中心とする主要メンバで検証した。

次年度では、がん領域のダミー経過記録を 1 万例ほど作成し、LLM を用いた医療データの抽出（がん領域 Json マスターでフィルタリング）を行い、Json データをデータベース化した。このデータベースの検索を行う簡易アプリケーションを試作し、臨床研究に資する十分な検索結果を得られることがわかった。

研究分担者氏名・所属研究機関名及び所属研究機関における職名

中島貴子・京都大学医学研究科・教授
松本繁巳・京都大学医学研究科・特定教授
黒田知宏・京都大学医学研究科・教授
吉原博幸・京都大学医学研究科・研究員（名誉教授）
小林慎治・岐阜大学大学院医学系研究科・特任講師
粂直人　・広島大学病院医療情報部・特定教授
横田理央・東京科学大学・学術国際情報センター・教授
加藤康之・新医療リアルワールドデータ研究機構株式会社・客員研究員
江口佳那・京都大学大学院情報学研究科・講師

Ａ．研究目的

本研究は、武藤らの癌化学療法支援システム（CyberOncology）を基盤に、千年カルテで蓄積した多施設電子カルテの非構造化テキストを活用し、LLM により医薬品の有効性・安全性評価に資するアウトカム抽出法を開発するものである。従来、治療効果や有害事象の情報は経過記録などに埋もれ、人手依存の処理が必要だったが、本研究はそれを自動化し、医薬品評価の効率化、リアルタイム監視、治療効果判定の迅速化、リスク管理精度の向上を目指す。併せて現行法制度下での実現可能性も検討し、医薬品開発と医療提供の高度化を通じて患者安全と福祉の向上に貢献する。

Ｂ．研究方法

本研究は、オープンソースの英語版 LLM を基盤とし、その後日本語化したモデルを出発点として、診療ガイドライン等の医学知識の学習、大規模電子カルテデータによる追加学習、さらに実臨床データを用いた構造化精度の評価へと段階的に発展させるものである（東京科学大学の Swallow)。ここでいう構造化精度とは、電子カルテ内に記載された非コード化テキスト情報から、診断、症状、治療内容、有害事象などの意味ある臨床情報を適切に抽出し、解析可能な形に整理する能力を指す。研究全体は 4 系統のタスクから構成され、年度ごとに明確な成果目標を設定して進める。LLM のファインチューニングは計 3 回を予定しており、第 1 回は診療ガイドライン等による知識学習、第 2・第 3 回は LDI が保有する大規模電子カルテデータによる学習を行う。構造化精度の到達目標は、第 1 回 90%以上、第 2 回 95%以上、第 3 回 98%以上である。

（倫理面への配慮）

Ｃ．研究結果

本研究の初期段階では、現行の法制度下において医療分野での LLM 開発および活用に伴う課題と技術的可能性を整理した。Llama3.3-Swallow-70B モデルを用い、千年カルテ由来の経過記録から初期学習モデルを構築。構造化データ抽出において、日時や治療歴、判定、誤記、Stage 分類、多言語対応といった項目を対象に新たな評価指標を提案し、実証を進めた。

Meta 社の Llama モデルをベースに量子化処理を施した結果、5〜6 ビットが精度と計算資源のバランスに優れていた。また、プロンプト表記との組み合わせ次第で量子化後の精度が変動するため、詳細な設計が精度維持に不可欠であることも明らかになった。

英語モデルの日本語応用に関しては、Llama3.3-Swallow-70B および Llama3-Preferred-MedSwallow-70B の検証結果から、日本語による継続学習は事象把握の曖昧

化を招く傾向があり、とくに「年」の省略表現によってタイムラインの混乱が起こることが示された。これはMeta社の多言語モデルにも共通する現象で、日本語の文法構造に起因する根本的な課題である可能性がある。
これまでのLLMファインチューニングの結果を踏まえて、ダミーデータではあるが、がん領域の経過記録を1万例ほど作成し、LLMを用いた医療データの抽出（がん領域Jsonマスターでフィルタリング）を行い、Jsonデータをデータベース化した。このデータベースの検索を行う簡易アプリケーションを試作し、臨床研究に資する十分な検索結果を得られることがわかった。最終年度では、これを用いた検索結果の評価を行う予定である。

構造化精度検証の自動化では、辞書構造・論理検証機能を備えた評価ツールの開発を進めており、LDIの大規模データにも対応できる基盤を構築中である。今後は医療用LLMの実運用を見据え、検証自動化の高度化が不可欠となる。

Ｄ．考察
本研究で得られた結果は、医療分野における日本語LLMの実用化に向けて、いくつかの重要な知見を提供するものである。
まず、量子化処理に関して、5〜6ビットが精度と計算資源のバランスにおいて最適であったという結果は、医療機関における限られた計算環境下でのLLM運用可能性を示唆している。ただし、量子化後の精度がプロンプト表記との組み合わせによって変動するという知見は、臨床現場での実装においてプロンプト設計の標準化が不可欠であることを意味する。しかしながらLLMの高精度化が著しい現段階での標準化には限界があるため、量子化後の精度検証は継続的な取り組みが必要である。
次に、英語モデルの日本語適用における課題として、日本語継続学習が事象把握の曖昧化を招く傾向が確認されたことは、特に注目すべき点である。「年」の省略表現によるタイムラインの混乱は、日本語の文法構造に起因する根本的な課題である可能性が示されたが、医療領域においては時系列情報の正確な把握が診療経過の理解や治療方針の決定に直結するため、この問題は極めて深刻である。本研究では、この課題に対してLMSYS-Chat-1Mの対話履歴を翻訳し、Llama 3.1 405B Instructによる日本語応答の自動生成とLlama 3.1 70Bによるスコアリング評価を組み合わせた新たなファインチューニング手法を導入することで、日本語指示への応答精度を大幅に改善することに成功した。この手法は、単純な日本語データによる継続学習ではなく、高品質な対話データの選別と多段階の評価プロセスを経ることで、日本語特有の曖昧性に起因する精度低下を効果的に抑制したものと考えられる。重複や冗長性のある学習データをフィルタリングする工程を組み込んだことも、学習効率と最終的な出力品質の向上

に寄与したと推察される。
また、がん領域の経過記録1万例からJsonデータを抽出・データベース化し、簡易アプリケーションによる検索で臨床研究に資する十分な結果が得られたことは、LLMを活用した医療データの構造化と二次利用の実現可能性を実証した点で意義が大きい。従来、非構造化テキストである経過記録から研究に必要な情報を抽出するには多大な人的労力を要していたが、本手法により、その過程の大幅な効率化が期待できる。ただし、現段階ではダミーデータによる検証であるため、実際の臨床データを用いた場合の精度や、抽出漏れ・誤抽出の頻度についてはさらなる評価が必要である。最終年度に予定している検索結果の評価により、実臨床データへの適用可能性がより明確になるものと期待される。
構造化精度検証の自動化に関しては、辞書構造・論理検証機能を備えた評価ツールの開発が進められており、大規模データへの対応基盤が構築されつつある。医療用LLMの実運用においては、出力結果の信頼性を継続的かつ効率的に担保する仕組みが不可欠であり、本ツールの高度化は品質保証の観点から極めて重要な取り組みである。
本研究は日本語医療LLMの構築において、英語事前学習モデルの日本語適用時に生じる固有の課題を明らかにするとともに、高品質データ選別に基づくファインチューニング手法によってその克服が可能であることを示した。さらに、構造化データ抽出からデータベース化・検索に至る一連のパイプラインの実現可能性を提示したことは、医療AIの実用化に向けた重要な前進と位置づけられる。

Ｅ．結論

英語で事前学習されたLLMを日本語に適用する場合、精度の低下が見られるが、これは日本語構文の特性や非明示的な時系列表現が要因である。これに対して、本研究では新たなファインチューニング手法を導入し、日本語指示への応答精度を大幅に向上させることに成功した。
その手法は、LMSYS-Chat-1Mの対話履歴を翻訳し、Llama 3.1 405B Instructを用いて日本語の応答文を自動生成するもの。続いて、Llama 3.1 70Bモデルによるスコアリング評価により最良の応答を選別する工程を組み込んだ。加えて、重複や冗長性のある指示文・応答文をフィルタリングし、学習データ全体の品質を高めた。
この一連の工程により、Llama3.3-Swallow-70Bにおいても日本語に特化したファインチューニングが可能であることが示され、医療領域での日本語LLM実装に向けた重要な成果を得た。
LLMを用いた経過記録からの医療データの抽出を行い、生成したJsonデータをデータベース化し、このデータベースの検索を行うアプリケーションを使い、臨床研究に資する十分な検索結果を得られることがわかった。

今後は、LLM の日本語継続学習時の性能劣化を抑制しつつ、多言語モデルの相互運用性を確保する設計原則の確立が求められる。本研究の成果は、日本語 LLM の基盤技術として国内外の医療 AI 研究にも貢献しうる。

Ｆ．健康危険情報

なし

Ｇ．研究発表

1.　論文発表

なし

2.　学会発表

[1] 吉原博幸，他，電子カルテ由来臨床データの知識ベース化：ナラティブデータからLLMを用いて臨床データを構造化する，第45回医療情報学連合大会，公募パネルディスカッション2，2-D-1，2025.

[2] 加藤 康之，他，LLMを用いた経過記録情報のユーザ開放型解析環境，第30回日本医療情報学会春季学術大会，0B4-01，2026.

Ｈ．知的財産権の出願・登録状況（予定を含む。）

1. 特許出願

名称：取得システム及び取得プログラム

特許内容：JSON マスタを用いた経過記録の構造化手段

発明者：加藤 康之 外２名

出願番号：特願２０２５－１３９３４５

特許出願人：新医療リアルワールドデータ研究機構株式会社

出願日：２０２５年 ８月２５日

2. 実用新案登録

なし

3. その他

なし