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（24AC0401）

研究代表者 京都大学医学研究科　教授・武藤 学

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する。さらに LLM の基本的な適用範囲と能力の評価も行う。

この段階では、現在、千年カルテプロジェクトで収集されている臨床情報データベース（経過記録等：3 億 9600 万文書、3TB）から非構造化情報を収集し、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(武藤、松本)を中心とする主要メンバで検証する。また、この年度の終わりには、初期の課題を特定し、次年度の改良に向けた基盤を築く。

研究分担者氏名・所属研究機関名及び所属機関における職名

松本 繁巳・京都大学医学研究科・特定教授
中島 貴子・京都大学医学研究科・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・広島大学病院インドネシア医療関連共同研究・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・シニアフェロー
江口 佳那・京都大学大学院情報学研究科・講師

Ａ．研究目的

治療効果の判定や有害事象に関わる情報は経過記録や報告書などの大量の非構造化テキストデータとして記録されているため、機械的に処理することが困難であり多くの人手を要している。近年、最先端の自然言語処理技術として大量のテキストデータを学習させた大規模言語モデル（LLM）が人間を上回る精度を示しつつある。

LLM の開発のためには大量のテキストデータと計算資源が必要となるが、我々は千年カルテプロジェクトで多施設から大量のテキストデータを含む電子カルテ情報を収集し、蓄積している。このテキストデータを利用して LLM を開発し、従来は手動でしか処理できなかった膨大なテキスト情報から、医薬品の安全性と有効性に関連する重要な知見を自動で抽出することが期待できる。これを実現することで、リアルタイムの医薬品監視、治療効果の迅速な評価、そしてリスク管理の精度向上に寄与し、最終的に医薬品開発と医療サービス提供の向上に資することを目指す。

Ｂ．研究方法

現時点利用可能な高性能なオープンソースの LLM（ベースモデル）を選定し、電子カルテの経過記録等（非構造化情報）を用いてベースモデルの構造化精度を検証する。ベースモデルとしては、本研究の分担研究者が開発する東京科学大学の Swallow-70B モデルを出発点とし、他のモデルとの性能比較等を行いながら以下の手順で研究を実施する。

１）構造化データ抽出機能の新たな評価方式の提案
２）LLM における構造化精度の量子化依存性の明確化
３）プロンプト表記の量子化依存性の明確化
４）高性能英語モデルの日本語化における課題の明確化
５）構造化精度検証の自動化

Ｃ．研究結果

・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。
・LLM の基本的な適用範囲と能力の評価を行った。
・この段階では、現在、千年カルテプロジェクトで収集されている臨床情報データベースの経過記録等から非構造化情報を収集し、初期の LLM モデルを構築してトレーニングを行なった。
・研究成果は以下の通りである。

１）構造化データ抽出機能の新たな評価方式の提案

実臨床データを基にした検証用ダミー資材を作成し、経過記録から抽出すべき臨床項目を特定した後、その正確性を測るものである。評価のポイントは以下の通りである：

a）日付の認識：西暦や日付の省略形を正確に年月日として認識できるか。
b）JSON 形式での抽出：項目キーと値の組み合わせを精緻に抽出する能力。
c）治療歴の認識：治療の開始日、終了日、治療ライン、薬剤名、投与量の認識度合い。
d）効果判定の認識：判定手段や判定内容の正確な認識。

e)誤字や Stage の間違いの指摘:誤字や Stage の間違いを指摘し、スコア化する。
f）多言語対応：英語・日本語問わず、意味が同じであれば正解とする。

2）LLM における構造化精度の量子化依存性の明確化

多くの分野で最も利用されている Meta 社の Llama モデルを対象に、量子化の依存性を明らかにした。電子カルテの構造化目的では、Meta-Llama-3-70B-Instruct において 5〜6 ビットの量子化が妥当な選択となることを明らかにした。量子化ビット数（Q2〜Q8）と構造化精度の関係を図 1 に示す。図の横軸は、必要とされる GPU メモリ量である。

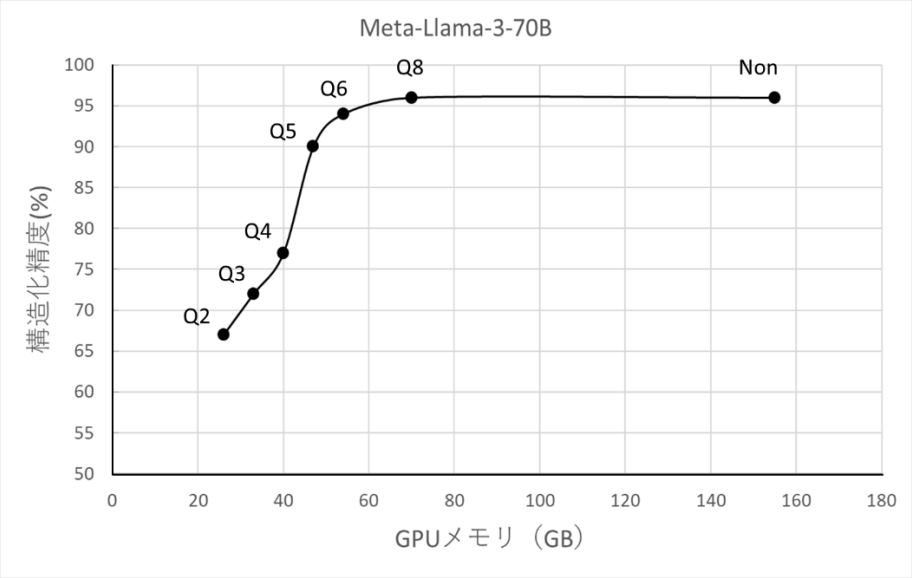
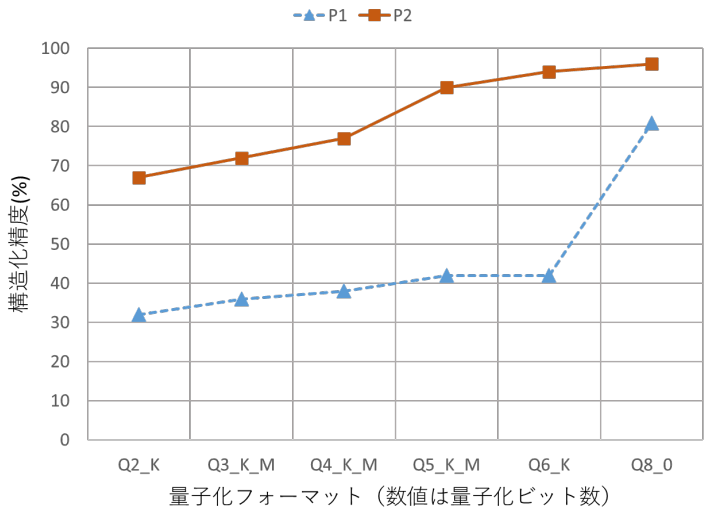


図 1. 必要とする GPU メモリと構造化精度の関係（Q2〜Q8:量子化ビット数）

3）プロンプト表記の量子化依存性の明確化

プロンプト表記の違いによる構造化精度の量子化依存性を明らかにした。5 ビット量子化の領域では、プロンプトの詳細な記述が求められることも明らかにした。



【簡易プロンプト例 P1】:
以下のテキストに書かれた事柄を時系列に JSON 形式で構造化し、可能な限り詳細に抽出してください。

【詳細プロンプト例 P2】:
以下のテキストに書かれた事柄を時系列に詳細に全ての事象を JSON 形式で構造化し、改行を加えて見やすく表示してください。なお症状や検査検体、診断名、TNM 分類、ステージ、転移部位、遺伝子変異などについても詳細に抽出してください。

図 2. プロンプトの違いによる構造化精度の比較（P1:簡易指定、P2:詳細指定）

4）高性能英語モデルの日本語化における課題の明確化

日本語ベンチマークで高いスコアを出している Llama-3-Swallow-70B-Instruct、さらにこれを医療領域にファインチューニングした Llama3-Preferred-MedSwallow-70B について、構造化精度の量子化依存性を詳細に検証した（図 3 及び図 4）。結果は、英語で高度にトレーニングされた LLM に対して日本語での継続学習を行うと、英語による構造化処理の性能に比べて、日本語の場合、その性能が低下することが明らかとなった。

これに対処するため英語環境の能力を維持しつつ、日本語指示に従う能力を獲得するようなトレーニング法の研究を行っている。

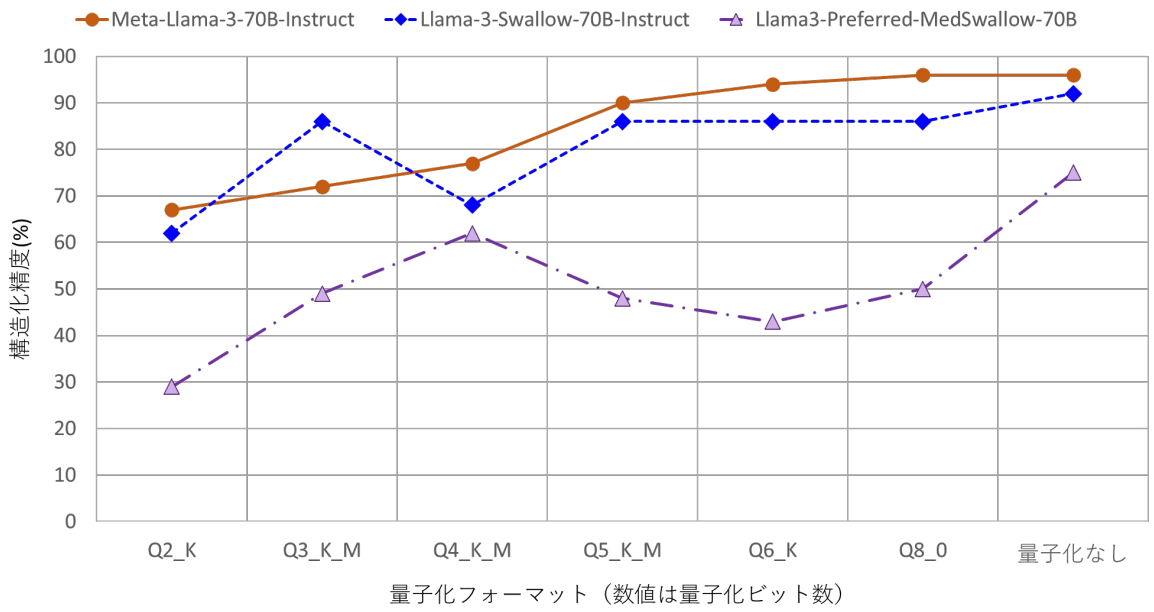


図3. 日本語継続事前学習モデルと構造化精度の関係

事象項番	経過記録	臨床項目	項目数	Meta-Llama-3-70B							Llama-3-Swallow-70B							Llama3-Preferred-MedSwallow-70B						
				Q2	Q3	Q4	Q5	Q6	Q8	なし	Q2	Q3	Q4	Q5	Q6	Q8	なし	Q2	Q3	Q4	Q5	Q6	Q8	なし
1	2019年12月初旬に胸痛、背部痛を自覚。近医を受信し左肺門部腫瘤を指摘され	診療日、症状	2	2	2	2	2	2	2	2	1	2	2	2	2	2	2	2	0	0	0	2	0	0
2	1月7日に当院を紹介受診。 同日に胸腔穿刺を施行し、得られた胸水から肺腺癌と診断 T4N1M1c StageⅣC, BRA, PUL, LYM, PLE, OSS, EGFR(L858R+), KRAS−, BRAF(V600E)−, ROS1−, PDL−1<0%, ALK−	診療日	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
		検体キー、検体名	2	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		診断キー、診断名	2	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	0	2	2
		TNMキー、TNM分類	2	1	2	2	2	2	2	2	1	2	2	2	2	2	2	1	0	0	0	1	0	2
		Stageキー、Stage	2	0	0	2	2	2	2	2	0	1	2	2	2	2	2	0	0	0	0	1	0	2
		指摘キー、誤字やStage間違いの指摘	3	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0
		転移キー、転移値（5つ）	6	6	0	6	6	6	6	6	0	4	6	6	0	0	6	0	4	3	1	5	4	6
		バイオマーカーキー（6つ）	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
		バイオマーカー値（6つ）	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
3	1/31～：C1-7頸椎転移に対し緩和的放射線治療(30Gy/10fr)施行	開始日、治療キー、治療内容、部位キー、部位名、線量キー、線量値	7	4	7	5	7	5	7	7	0	5	5	5	7	7	7	2	0	7	4	0	0	7
4	2/6～：1stline Osimertinib 80mg/dayを開始（28日分）。	開始日、ラインキー、治療ライン、治療キー、治療タイプ、薬剤キー、薬剤名、投与キー、投与量、期間キー、投与期間	11	7	9	7	10	11	11	11	10	11	9	10	10	11	11	3	8	9	10	6	9	9
5	11/4：CTにて左肺門部腫瘤、左肺尖結節、肝転移、骨転移増大を認めPDと判断。	診療日、判定手段、検査内容、判定内容	4	1	0	3	4	4	4	4	3	3	3	3	4	4	4	0	0	2	2	3	0	4
6	11/19～2/10：2nd line CBDCA/ PTX/ Bev/ Atezoを4course施行	開始日、終了日、ラインキー、治療ライン、治療キー、治療タイプ、薬剤キー、薬剤名、投与キー、投与量	10	6	7	6	8	10	10	10	8	9	7	9	9	10	10	2	8	8	9	2	8	8
7	2021/3/2：効果判定にて原疾患の増悪を認めPDと判断。	診療日、判定手段、検査内容、判定内容	4	3	4	4	4	4	4	4	3	4	3	3	4	4	4	2	0	2	2	3	0	4
8	3/16～4/13：3rd line DOC+RAM 2course施行	開始日、終了日、ラインキー、治療ライン、治療キー、治療タイプ、薬剤キー、薬剤名、投与キー、投与量	10	7	7	6	8	10	10	10	8	9	7	9	10	10	10	2	8	8	9	4	8	8
9	5/18：効果判定にて両肺の小結節は増加・増大と両測胸水の増加がありPDと判断	診療日、判定手段、検査内容、判定内容	4	3	4	4	4	4	4	4	3	3	3	3	4	4	4	2	0	2	2	3	0	4
10	5/20：胸水コントロール目的に入院。	診療日、治療キー、治療内容	3	2	3	3	3	3	3	3	3	3	0	3	3	0	0	0	0	0	0	0	0	0
11	5/21：左癌性に対して胸水左胸腔ドレーン挿入	診療日、治療キー、治療内容	3	2	3	3	3	3	3	3	3	3	0	3	3	3	3	0	0	0	0	0	0	0
12	5/24：左胸膜癒着術(ユニタルク4g)を施行	診療日、治療キー、治療内容	3	2	3	3	3	3	3	3	3	3	0	3	3	3	3	0	0	0	0	0	0	0
13	5/27～：4th line EGFR-TKI rechallenge (Afatinib 20mg/day)開始	開始日、ラインキー、治療ライン、治療キー、治療タイプ、薬剤キー、薬剤名、投与キー、投与量	9	7	7	5	8	9	9	9	8	9	6	8	8	9	9	4	7	7	7	7	7	7
正解項目数の合計（＝構造化精度%）			100	67	72	77	90	94	96	96	69	86	70	86	86	86	92	35	49	62	60	49	50	75

図4. 日本語継続追加学習モデルと構造化精度の詳細

５）構造化精度検証の自動化

上記１）の評価方式を用いてLLMの構造化精度を判定するのに人手で行って来たが、これをプログラムにより自動判定するアプリケーションの開発に着手している。開発は、以下のステップで進める予定であり、①のステップを完了している。

① ステップ1：上記１）の方式に特化した辞書構造と論理チェックをプログラムで実現
② ステップ2：LLMの構造化を一定の形式に固定化する方式の実現
③ ステップ3：今後学習予定の大規模なLDIデータに柔軟に対応できる 辞書構造と論理チェックを可能にする方式の実現

Ｄ．考察

英語圏で学習された高性能なモデルにおいて日本語継続学習を行うと事象の把握が曖昧のなる傾向が検証された。すなわち経過記録として時系列的に記載されたカルテにおいて、よく散見される年月日の「年」の省略がLLMにおいて混乱を来す現象である。この現象は、ベースモデルを開発した会社（Meta 社）が独自に行っている多言語化においても同様の現象が現れることから日本語の構造が持つ特有の本質的な何かを明らかにした可能性がある。カルテ記録から臨床情報を抽出する上で事象把握の厳密性は極めて重要であり、英語圏での性能を支持しながら日本語継続学習を行うノウハウの確立が今後重要である。

Ｅ．結論

英語ベースでトレーニングを行なったLLMを出発点としているが､日本語学習の結果、LLMの性能が低下することがわかった。ただし、これを改善するために指示に従う能力を向上するようなファインチューニング（後述）を行った結果、オープンなモデルでは最高性能を達成することができた。

新規ファインチューニングの方法：

人間と大規模言語モデルの対話履歴を収録した LMSYS-Chat-1M データセットの指示文を邦訳し、オープンなモデルの中でトップクラスの対話能力を有する Llama 3.1 405B Instruct を用いて応答文を自動生成した。また、Llama 3.1 構築の方法論に倣い、複数の応答文を生成してから Llama 3.1 70B Instruct に選好を自動採点させ、最良の応答文を選択するという工夫を取り入れた。さらに、重複する指示文や機械的な指示文、無用な繰り返しを含む応答を検出・削除することで、データの品質を向上させた。

ChatGPT などのクローズドなモデルは性能は高いが、OpenAI のサーバにデータを転送する必要があるため､医療データに用いることは困難である。この点で院内からデータを出すことなく使うことができるオープンなモデルで最高性能を達成できたことは意義が大きい。

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

石田茂樹、 加藤康之、横田理央、「大規模言語モデルを用いたカルテ記載情報の構造化精度自動評価手法の検討」、第 47 回計算数理工学フォーラム（The 47th JASCOME Forum）、2025 年 3 月 21 日.

Ｈ．知的財産権の出願・登録状況（予定を含む。）

なし