TSUBAME 共同利用　令和 7 年度　学術利用　成果報告書

利用課題名　**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**

英文：Research to establish a methodology for extracting outcomes for evaluating the efficacy and safety of pharmaceuticals and other products using Large Language Models (LLM)

利用課題責任者

武藤 学 / Manabu Muto

所属

国立大学法人 京都大学 / Kyoto University

邦文抄録（300 字程度）

本研究では、大規模言語モデル（LLM）を用い、多施設電子カルテの非構造化テキストから医薬品等の有効性・安全性評価に資するアウトカムを抽出する方法論を検討した。TSUBAME の GPU 資源により、Swallow/Qwen/GPT-OSS 系モデルの継続事前学習・SFT・推論評価、医療ベンチマーク、自然言語から SQL を生成するデータベース解析フローを実装した。約 7,000〜8,000 症例規模のデータ登録、SQL 生成・実行、薬剤×効果判定の可視化を行い、モデル性能の改善点と課題を整理した。

英文抄録（100 words 程度）

This project investigated a methodology for extracting clinically meaningful outcomes from unstructured electronic health record text using large language models. Using TSUBAME GPU resources, we prepared and evaluated medical-domain LLMs based on Swallow, Qwen, and GPT-OSS, including continued pretraining, supervised fine-tuning, inference pipelines, and benchmark evaluation. We also developed a natural-language-to-SQL workflow for clinical databases and demonstrated registration of approximately 7,000 to 8,000 cases, SQL generation and execution, and visualization of drug-by-outcome patterns. The results clarified both the potential of LLM-assisted outcome extraction and remaining issues in data mixture, evaluation, and clinical validation.

Keywords: Large Language Model, Electronic Health Record, Outcome Extraction, Medical Text Mining, TSUBAME/GPU

背景と目的

医薬品等の有効性・安全性評価では、診療録、検査レポート、処方・注射オーダー、退院サマリなど、電子カルテ内に蓄積された非構造化テキストをどのように再利用するかが重要である。一方で、臨床文書は表記揺れ、時系列の複雑さ、疾患名・薬剤名・検査値の関係、施設差、個人情報保護といった制約を含むため、従来のルールベース抽出や小規模モデルだけでは汎用的なアウトカム抽出が難しい。

申請書では、LLM を用いて多施設から収集した電子カルテ内の非構造化テキスト情報から、医薬品等の有効性と安全性評価に資するアウトカムを抽出する方法論を確立することを目的とした。令和 7 年度は、TSUBAME の GPU 資源を用いて医療用 LLM の継続事前学習・SFT・推論評価を試行し、臨床データベースを自然言語で解析するための基盤と評価方法を整備した。

本プロジェクトでは、非構造化カルテを構造化データへ変換し、自然言語問い合わせから SQL を生成してアウトカム集計・可視化へ接続する課題を、LLM と大規模 GPU 計算により解決することを目指した。その結果、医療ベンチマークに基づくモデル性能の把握、データ混合比・評価タスクの課題抽出、カルテデータを用いたデータベース解析フローのプロトタイプを得た。

概要

申請時の利用課題概要および計算資源計画は以下のとおりである。研究分野は「機械学習・深層学習系」、主利用ソフトウェアは PyTorch とし、70B パラメータ級モデルを 100B トークン規模で学習することを想定して 30 口の利用を申請した。共有ストレージは HDD 100TB、SSD 3TB を想定し、医療文書、評価データ、学習ログ、モデルチェックポイントの管理に充てる設計とした。研究実施は、(1) 医療文書・外部文献・OCR データの整備、(2) Swallow/Qwen/GPT-OSS 系モデルを用いた継続事前学習・SFT・推論評価、(3) IgakuQA、JMMLU_Medical、MMLU_Medical_JP、MedMCQA_JP、ACI-Bench、JMED-LLM、HealthBench 等による評価、(4) 自然言語から SQL を生成するデータベース解析フローの実装、の四つを中心に進めた。

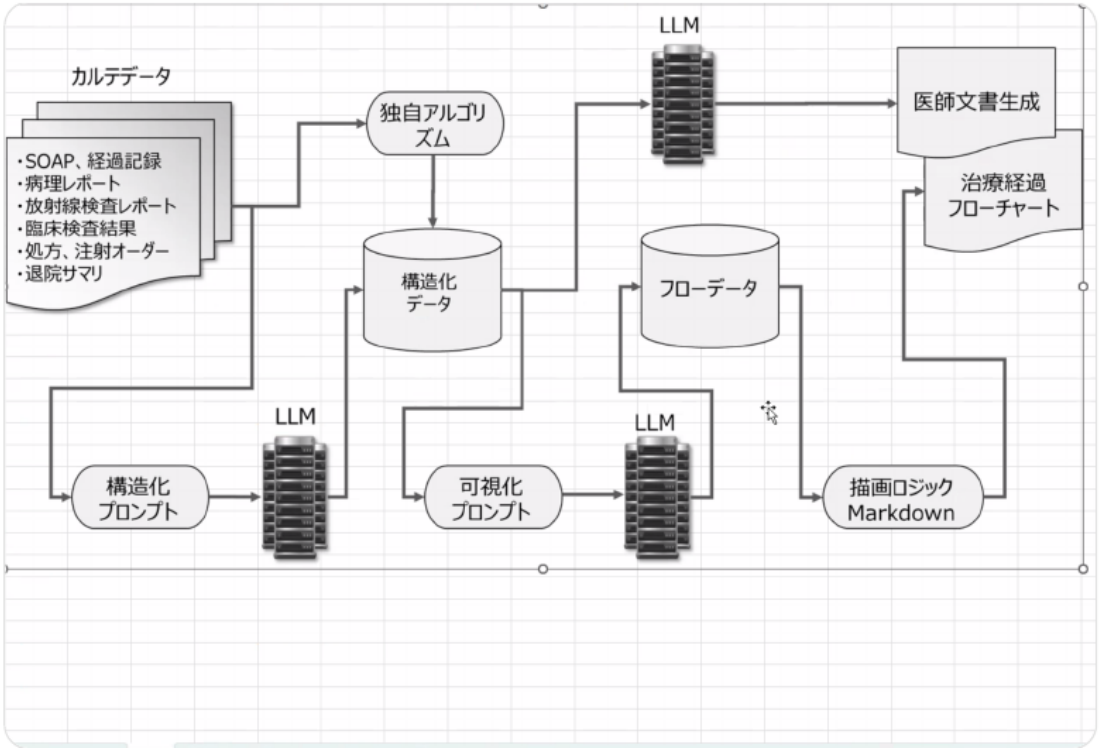

図 1　カルテデータから構造化データ・フローデータ・医師文書生成へ接続する LLM ワークフロー

図 1 のように、SOAP、経過記録、病理・放射線検査レポート、処方・注射オーダー、退院サマリ等を、構造化プロンプト、可視化プロンプト、独自アルゴリズム、LLM に接続し、治療経過フローチャートや医師文書の生成へ展開する設計を検討した。これは申請書に記載した「電子カルテの非構造化情報を活用し、医薬品評価プロセスを効率化する」という目的に対応する。

結果および考察

1. 医療用 LLM の学習・評価基盤

議事録では、医療データを混合した継続事前学習・SFT の候補として Qwen-3-8B、Qwen-3-30B-A3B、Qwen-3-32B、GPT-OSS-20B、GPT-OSS-120B 等が整理されている。TSUBAME 16 node を用いた学習時間の見積りとして、Qwen-3-32B が約 9.84 日、Qwen-3-30B-A3B が約 4.4 日、GPT-OSS-120B が約 15.2 日と記録されており、70B〜120B 級モデルを扱うには申請時の 30 口規模が妥当であることが確認された。

評価面では、医療 QA だけでなく、要約・構造化・医学知識・一般能力の複数タスクを同時に見る必要があることが明確になった。表 1 は、議事録に記録された Swallow 系 8B モデルの医療ベンチマーク結果を整理したものである。QA タスクでは学習後に必ずしも単調な改善は見られない一方、JMMLU_Medical、MMLU_Medical_JP、ACI-Bench の一部指標では改善が確認された。これは、医療ドメインへの適応が「医学知識を増やす」だけでなく、抽出・要約・構造化など利用目的に応じた評価設計を必要とすることを示す。

Model	IgakuQA	IgakuQA_Kinki	JJSIMQA	JMMLU_Medical	MMLU_Medical_JP	MedMCQA_JP	ACI BERT-F	ROUGE-1	ROUGE-2
8B-v0.2-Base	0.6732	17	0.6264	0.6912	0.7165	0.5419	0.6794	0.3750	0.1215
8B-Instruct-v0.2	0.6647	15	0.5846	0.6912	0.7186	0.5067	0.6882	0.4094	0.1267
8B-RL-v0.2	0.6591	16	0.5890	0.7022	0.7224	0.5096	0.6916	0.3959	0.1313

表 1　医療ベンチマークおよび ACI-Bench の代表値

20B トークン時点の Swallow evaluation では、LLM 分類器のデータ比率を 10%、30%、50%、70%、100% で比較し、平均スコアと数学・コードタスクへの影響を確認した。日本語・医療タスクでは一定の改善余地がある一方で、数学・コードなどの一般能力には低下が生じうるため、医療データ、PubMed/PMC、QA、common crawl、数理・コード系データの混合比を調整する必要がある。

2. 自然言語から SQL を生成するアウトカム抽出フロー

医薬品等の有効性・安全性評価に資するアウトカム抽出では、臨床文書から表記揺れを吸収した構造化データを作成し、自然言語の問いを SQL に変換して集計する仕組みが中心となる。議事録では、Swallow-70B による構造化データ作成、GPT-OSS-120B による SQL 生成、PostgreSQL による結果出力という役割分担が設計されている。

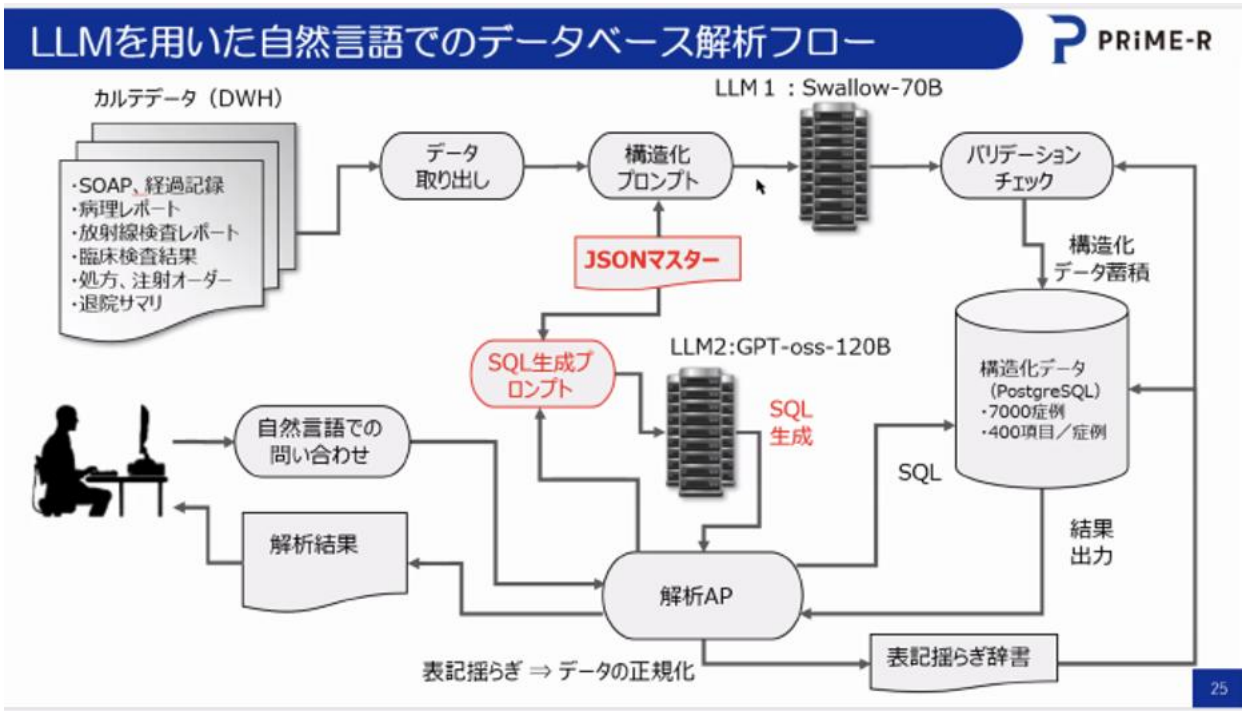

図 2　自然言語問い合わせから SQL 生成・解析結果出力に至るデータベース解析フロー

プロトタイプでは、約 7,000〜8,000 症例規模のデータをデータベースへ登録し、SQL の生成に約 12 秒、実行に約 1.5 秒程度という記録が得られている。これにより、電子カルテから抽出した構造化情報を対話的に検索し、薬剤名、効果判定、副作用、時系列イベントを横断的に集計する基盤の有効性が確認された。

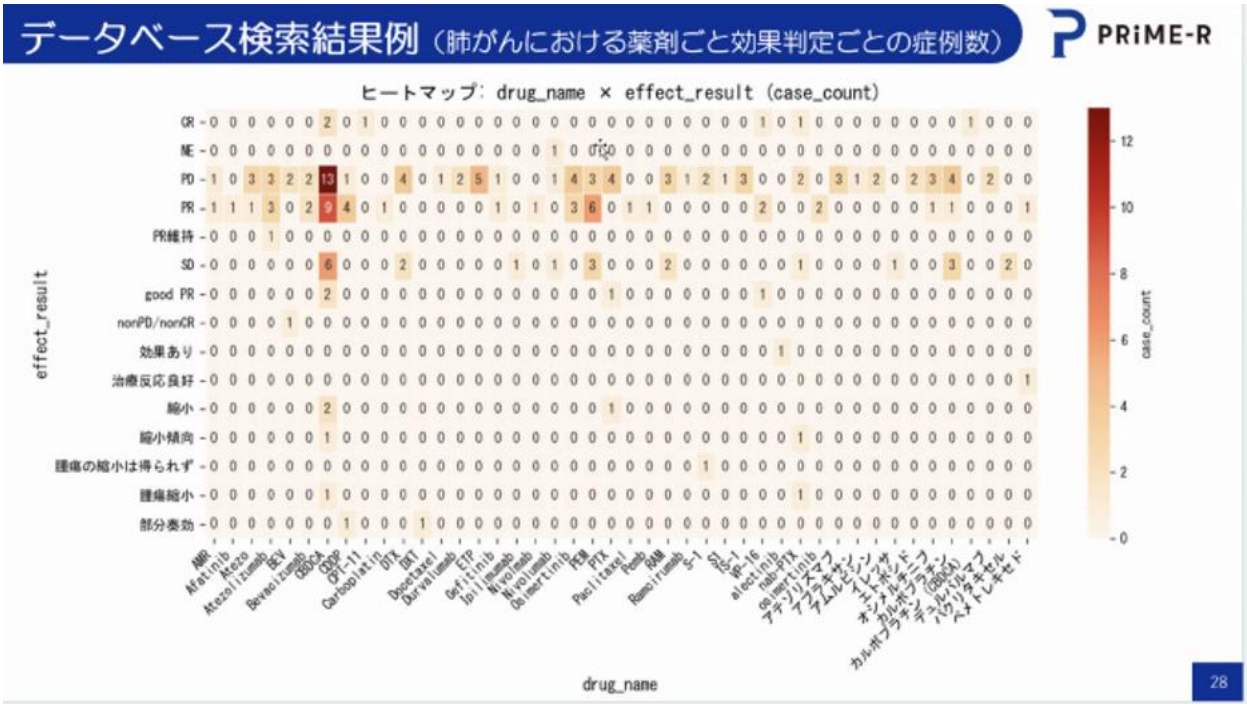

図 3　薬剤名×効果判定の症例数ヒートマップとして出力したデータベース検索結果

図 3 のような薬剤×アウトカムのヒートマップは、患者集団内で薬剤使用と効果判定の分布を概観する初期解析に有用である。最終的な薬効・安全性評価には交絡調整、時系列の解釈、診療科・疾患背景の層別化が必要であるが、LLM による構造化と SQL 生成を組み合わせることで、従来は手作業で時間を要していた候補抽出と仮説生成を高速化できる可能性が示された。

3. 構造化出力と評価自動化

LLM による抽出結果を実運用に近づけるには、出力の正確性を継続的に検証する仕組みが不可欠である。議事録では、実臨床を模した経過記録データを入力し、Swallow-70B などが JSON 形式で構造化出力を行い、高性能 LLM を用いて構造化精度を 1〜5 段階で評価する自動化方式が検討されている。

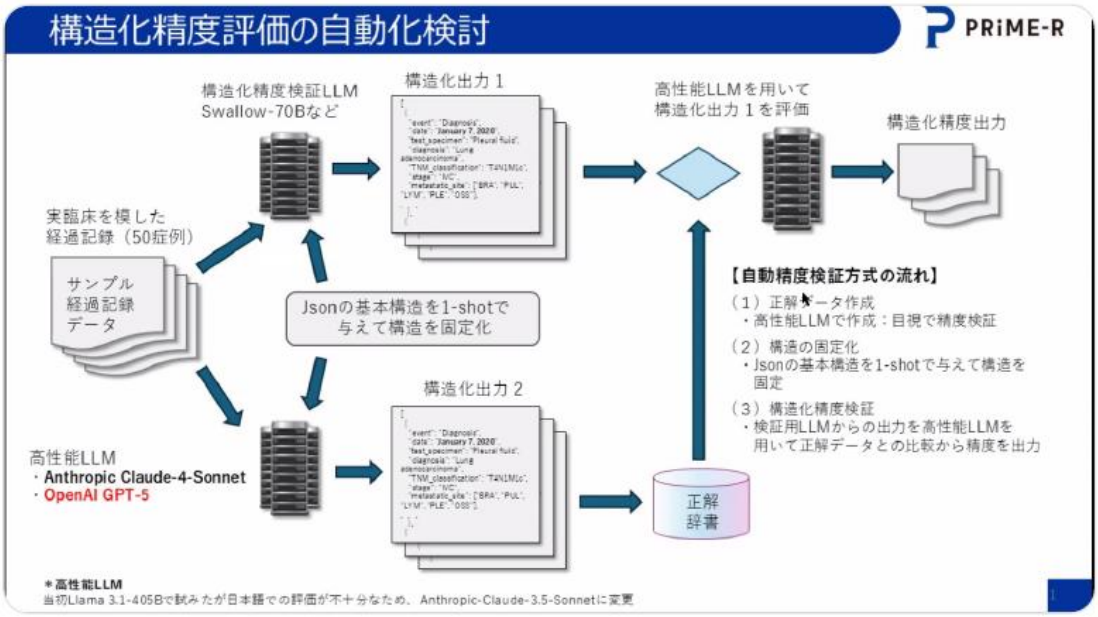

図 4　構造化精度評価の自動化

また、JMED-LLM の MRNER-disease、RRTNM、HealthBench、CareQA 等を候補評価セットとして調査した。MRNER-disease では partial F1、exact F1、RRTNM では TNM ステージ分類の accuracy を用いるなど、抽出タスクごとに指標が異なる。議事録では README と再実行結果の乖離が記録されており、評価データの生成方法、正解ラベル、プロンプト、量子化条件を再点検する必要が明確になった。

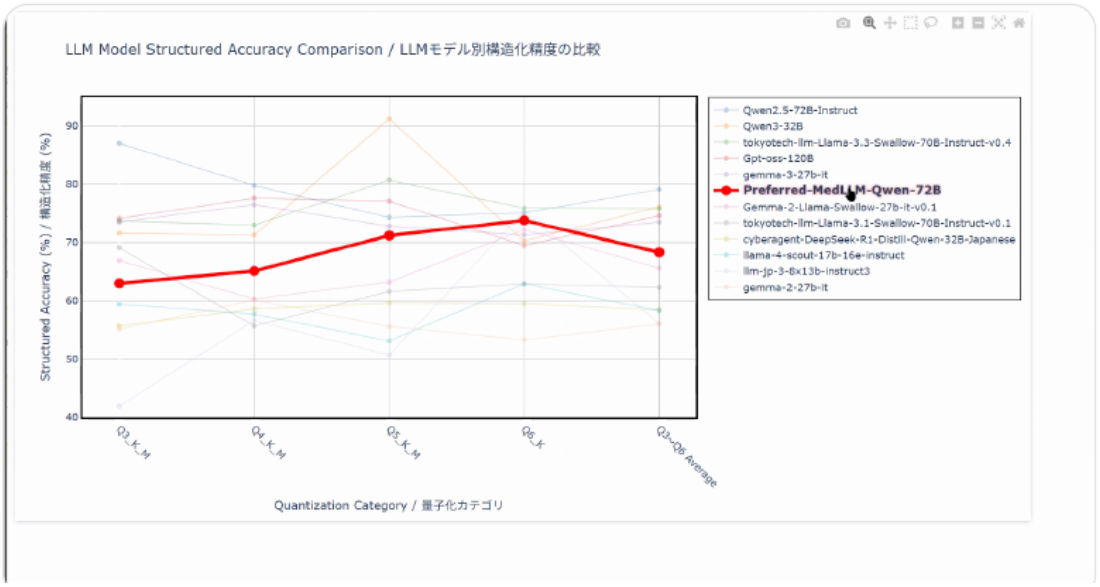

図 5　複数モデル・量子化カテゴリにおける構造化精度比較

4. データ整備・運用上の知見

医療文書の学習・評価では、病院内部データだけでなく、PubMed/PMC、J-STAGE OCR、医療ガイドライン、QA 形式データ、臨床試験や診療記録を模したデータ等を組み合わせる必要がある。議事録では、ABEJA OCR や fastText を用いた PDF/OCR データ処理、JSON データをデータベースに保存するシステム、表記揺れ辞書、構造化プロンプトの改善が進められている。

運用面では、vLLM の reasoning parser、LM Studio/Ollama、LiteLLM、DGX-B200 などの検証も記録されている。LLM 推論は研究用環境であってもセキュリティ、ネットワーク、リモートアクセス、ログ管理が重要であり、特に医療データを扱う際には、ホワイトハッカーによる安全性評価、RAG/ガードレール、個人情報を露出させない出力制御が今後の実装要件となる。

観点	令和 7 年度の成果	今後の課題
モデル開発	Swallow/Qwen/GPT-OSS 系モデルの候補整理、継続事前学習・SFT・RL 評価の試行、20B トークン時点評価。	医療データと一般能力データの混合比最適化、70B〜120B 級モデルの再現性ある学習。
評価	IgakuQA、JMMLU_Medical、MMLU_Medical_JP、MedMCQA_JP、ACI-Bench、JMED-LLM 等で比較。	評価データの正解性、プロンプト、量子化条件、LLM-as-a-Judge の妥当性検証。
アウトカム抽出	自然言語から SQL を生成し、薬剤×効果判定のヒートマップを出力するプロトタイプを構築。	交絡調整、時系列解析、臨床医レビュー、多施設データへの一般化。
データ基盤	OCR、JSON、PostgreSQL、表記揺れ辞書、構造化プロンプトの整備。	個人情報保護、アクセス制御、データ品質、施設差の標準化。

表 2　令和 7 年度の成果と今後の課題の整理

まとめと今後の課題

本年度は、申請書で掲げた「LLM を用い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方法論」の実現に向け、TSUBAME の大規模 GPU 資源を前提としたモデル学習・推論・評価・データベース解析の一連の基盤を整備した。具体的には、医療用 LLM の候補モデルと学習計画を整理し、医療ベンチマークおよび構造化評価で性能を確認した。また、カルテデータを構造化し、自然言語問い合わせから SQL を生成して薬剤×アウトカムを集計・可視化するプロトタイプを構築した。

今後は、第一に、評価の再現性を高めるため、データセット、プロンプト、量子化条件、採点基準を固定し、臨床医によるラベル検証を行う。第二に、モデル学習では、医療文書を入れることで医学タスクが向上する一方、数学・コードなど一般能力が低下する可能性があるため、データ混合比と SFT/RL の設計を継続的に改善する。第三に、アウトカム抽出では、単純な症例数集計から、時間軸、背景因子、交絡、施設差を考慮した解析へ発展させる。第四に、実運用を視野に、セキュリティ、RAG、ガードレール、ログ管理、個人情報保護の要件を組み込む。

これらにより、電子カルテに含まれる非構造化情報を安全かつ再現性高く活用し、医薬品評価プロセスの効率化、臨床研究の迅速化、患者安全と医療サービスの質向上につながる基盤技術として発展させる。