NEDO（国立研究開発法人新エネルギー・産業技術総合開発機構）
さくらインターネット株式会社
国立大学法人東京大学
株式会社ABEJA
国立研究開発法人理化学研究所
学校法人国際医療福祉大学
学校法人藤田学園　藤田医科大学
国立大学法人東京科学大学
国立大学法人九州大学
株式会社ヘリオス

医療現場の事務作業を支援する高性能な日本語 LLM を開発しました
―日本の医療特性を踏まえた安全性検証により医療 LLM の社会実装を後押し―

NEDOが推進した「AIの安全性確保に関する研究開発・検証等の推進事業／日本語版医療特化型LLMの社会実装に向けた安全性検証・実証」（以下、本事業）において、連名機関10者は、医療機関のオンプレミス環境または医療機関が管理する国内クラウド環境などの患者情報を安全に管理できる環境で運用可能で、世界最先端の商用LLMに迫る性能を有する医療業務支援向け日本語LLMを開発しました。

独自に構築したベンチマークによる検証の結果、専門医試験を模した学術試験において最大90.8%の正答率を達成し、比較対象とした主要な商用LLM（91.4%）に迫る水準に到達しました。あわせて、日本の医療特性を踏まえた安全性検証を実施し、医療現場での利用に求められる性能と安全性の両立を確認しました。

本研究開発の成果については、医療現場の業務効率化および医療の質向上に資することを目指し、今後、段階的に社会実装を進めていく予定です。

1. 背景

医療機関がAIを活用するにあたっては、以下の三つの構造的課題が存在します。

1）患者情報の管理に関する課題：一般的なAIサービスの多くは、患者情報が国外のサーバや外部事業者の管理下で処理される構造になっており、医療機関側で患者情報の所在や取り扱いを十分に把握・管理することが困難です。

2）データ標準化の課題：医療機関ごとに用語やコード体系が異なり、データの相互運用性が十分に確保されていません。

3）安全性基準の課題：医療現場におけるLLM※1活用にあたっての安全性基準が未整備であり、導入判断のよりどころが乏しい状況にあります。

本事業※2では、患者情報を安全に管理できる環境で運用可能であり、かつ主要なAIに匹敵する性能を有するAIの開発を目標として、（1）LLM開発、（2）安全性検証、（3）ユースケース検証の3点に取り組みまし

た。

2. 今回の成果

(1)患者情報を安全に管理できる環境で運用可能な高性能日本語LLMの開発

公開されているオープンなLLMをベースモデルとして、日本の診療ガイドライン・専門医試験問題・臨床事例などの医療分野の教材から生成したデータを学習させた追加学習※3モデルを開発しました。その結果、患者情報を安全に管理できる環境で運用可能でありながら、主要な商用LLMに迫る性能を実現しました。

主な成果として、専門医試験を模した学術試験において、外部文書を参照しながら回答する方式(RAG)を用いることで最大90.8%の正答率に到達し、比較対象とした主要な商用LLM(91.4%)に迫る水準に到達しました。また、日本の診療ガイドラインに沿った応答ができるかを評価する指標では、ベースモデルと比較して最大10.8ポイントの性能向上を確認しました(表1)。

加えて、独自アーキテクチャによる国産のフルスクラッチ開発※4モデルも構築しました。同規模のオープンモデルと比較して競争力のある性能を示し、将来の国産基盤モデル開発に向けた技術的知見を蓄積しました。

表1　代表的な追加学習モデルおよびフルスクラッチモデルの性能比較

カテゴリ	モデル	パラメータ	専門医試験	専門医試験（RAG）	ガイドライン
商用LLM（参考）	Claude Opus 4.5	—	90.6%	**91.4%**	69.8%
	GPT-5.2	—	91.2%	—	67.8%
	Gemini 3 Pro Preview	—	92.5%	93.3%	68.0%
追加学習モデル(東京大学開発)	**Weblab-MedLLM-GLM-4.7**	**355B MoE**	**83.5%（+1.6）**	**90.8%**	**63.7%（+10.8）**
	Weblab-MedLLM-Qwen3-235B-Thinking	235B MoE	82.7%（+1.7）	89.1%	60.6%（+4.1）
	Weblab-MedLLM-Qwen3-235B-Instruct	235B MoE	79.6%（+1.3）	87.6%	59.5%（+1.4）
	Weblab-MedLLM-gpt-oss-120b	120B MoE	81.1%（+5.2）	88.6%	45.6%（−7.7）
追加学習モデル（東京科学大学開発）	Medical-Qwen3-Swallow-30B-A3B	30B MoE	68.9%	—	—
	Medical-Qwen3-Swallow-32B	32B	58.9%	—	—
	Medical-Qwen3-Swallow-8B	8B	53.5%	—	—
	Medical-GPT-OSS-Swallow-120B	120B MoE	74.6%		
フルスクラッチ開発モデル（東京大学）	AscleLM1-30B-A10B	30B MoE	65.3%	—	—
	AscleLM1-10B	10B	41.5%	—	45.6%

・本資料に記載されている商用LLMの名称は、各社の商標または登録商標です

・東大・追加学習モデルの括弧内数値はベースモデル(追加学習前)からの変化幅

・専門医試験:眼科・外科・救急科・麻酔科・内科・産婦人科・整形外科の専門医試験過去問から作成した独自ベンチマーク

・専門医試験(RAG):上記の専門医試験を、診療ガイドライン等の参考文献をRAG(Retrieval-Augmented Generation: 検索拡張生成)で付与した条件で評価。最新の医学情報や出典に基づいた回答が求められる実運用に近い条件

・ガイドライン:複数の診療科の診療ガイドラインから作成した独自ベンチマーク。日本の医療現場の対応方針に即した回答が得られるかを評価

(2)日本の医療特性を踏まえた独自の安全性検証

LLMが医療情報を扱ううえで重要となる安全性の検証として、以下の多面的な取り組みを実施しました。

1）学習データに含まれる患者情報がLLMに記憶されるリスクを定量的に評価する手法の確立
2）患者情報を自動で検出・マスキングする機能の実装
3）日本の医療特性を踏まえた対話型安全性ベンチマーク（5万件超）の策定・公開およびモデル評価（図1）
4）攻撃耐性を評価する試験（6000件規模のレッドチーミング※5）の実施（図2）

検証の結果、追加学習を行った後もベースモデルと同等の高い安全性を維持できることを確認しました。一方で、ベースとなるLLMの選択が安全性維持を大きく左右することも明らかになり、医療分野でより安全なAIを開発する際の重要な知見を得ました。

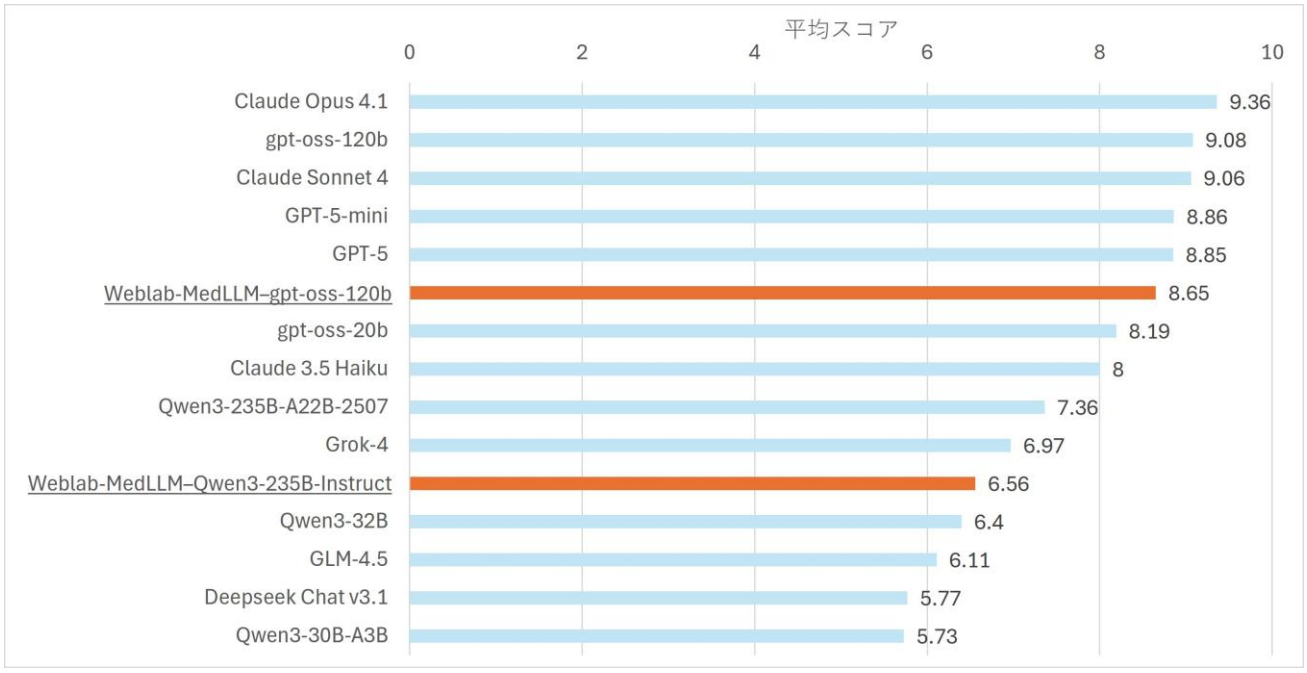

図1　対話型安全性ベンチマーク評価結果（抜粋）

・下線ありが本事業の成果

・スコアは10点満点で高いほど安全

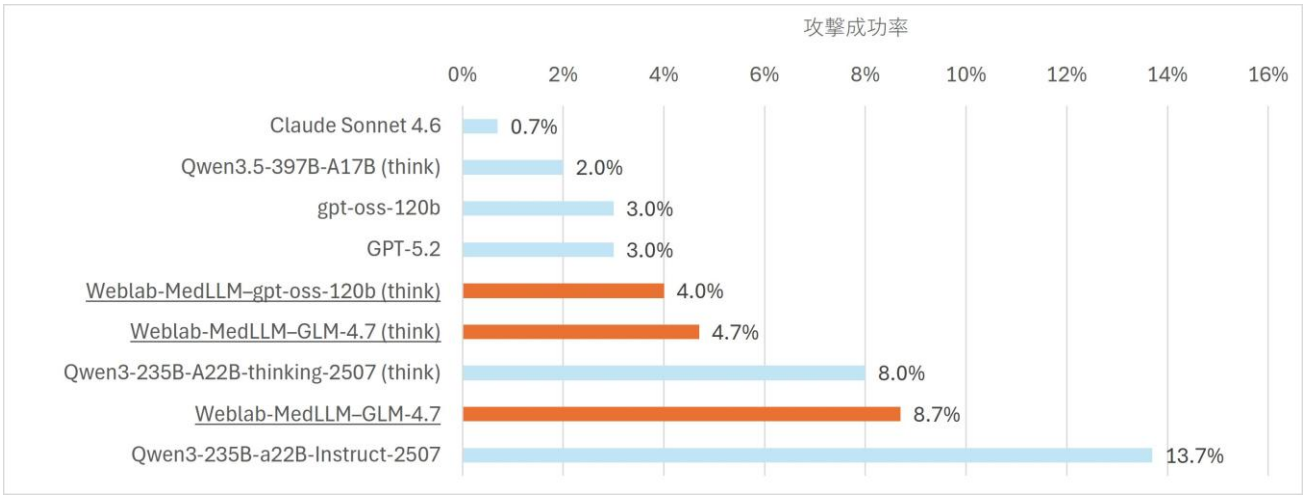

図2　モデル別の攻撃耐性（抜粋）

・下線ありが本事業の成果

・判定基準は攻撃成功率0%＝合格、0%超～20%未満＝要改善、20%以上＝不合格。(think)はreasoningを有効化していることを示す

(3)医療業務支援を想定したユースケース検証

医療従事者の事務的・文書的業務を支援することを目的として、以下の複数のユースケースにおいて技術的な実現可能性を検証しました。

1)検査名称からJLAC11コードへの自動変換:3医療機関のマスタデータで検証し、最大80.3%の精度を達成

2)症例データの自動整理(脳卒中レジストリ構築):人間の作業精度(94～95%)に対し、LLMで92.2%を記録

3)退院時サマリーの下書き作成:専門医9名による品質評価で、本事業の追加学習モデルが商用LLM相当の品質(5点満点で4.748、GPT-5.2比 -0.06ポイント)を達成

4)電子カルテへの自然言語による問い合わせ:複数の電子カルテシステムとの接続方法を確立し、自然言語による問い合わせが可能であることを確認

これらはいずれも医療従事者の事務作業・文書作成を補助するものであり、疾病の診断・治療そのものを行うものではありません。最終的な判断は医師および医療従事者が行います。

3. 今後の予定

本研究開発で得られた医療業務支援向けLLMは、医療現場の業務効率化および医療の質向上に資することを目指し、今後、関係機関と連携しながら段階的に社会実装を進めていく予定です。

社会実装にあたっては、安全性・信頼性の確保を最優先に取り組むとともに、医療機関をはじめとする関係機関との丁寧な対話を重ねながら進めてまいります。

【注釈】

※1 LLM

大規模言語モデル(Large Language Model)のことです。文章の生成・要約等を行うAI技術です。

※2 本事業

事 業 名 :AIの安全性確保に関する研究開発・検証等の推進事業

事業期間:2025年度

事業概要:AIの安全性確保に関する研究開発・検証等の推進事業

https://www.nedo.go.jp/activities/ZZJP_100327.html

※3 追加学習

既存のLLMに特定分野のデータを追加で学習させ、当該分野に特化させる手法のことです。

※4 フルスクラッチ開発

既存モデルを基にせず、設計から学習までを一から行うLLM開発手法のことです。

※5 レッドチーミング

攻撃者視点で意図的に攻撃を仕掛け、システムの脆弱(ぜいじゃく)性を体系的に評価する手法のことです。