UTokyo

PRESS RELEASE

東京大学大学院工学系研究科

東京大学 松尾・岩澤研究室、医療業務支援向け日本語 LLM および安全性検証ツール群を公開

――医療現場の DX 推進に向け、研究者・開発者が試し・活用できる主要成果物を公開――

発表のポイント

- ◆ 2026 年 5 月 28 日に公表した「医療現場の事務作業を支援する高性能な日本語 LLM」の研究開発成果（NEDO 事業）に基づき、東京大学 松尾・岩澤研究室が中心となって開発した医療特化型 LLM の追加学習モデル 4 種と国産フルスクラッチ開発モデル「AscleLM-1-10B」を Hugging Face 上で公開。
- ◆ 5 月 28 日公表の研究開発で実際に用いられた「レッドチーミング・アプリケーションおよびレポート」「PII（個人情報）検出・ルーティングプログラム」の 2 つの安全性検証ツールを、再現性確保および医療 LLM の安全な活用支援を目的に公開。
- ◆ モデルと安全性検証ツールを公開することで、研究者・開発者が開発成果を実際に試し、その有用性を確かめたうえで自身の研究・サービスや医療現場の DX に活用・導入できる。これにより、知見の還元と医療業務支援 LLM の安全な社会実装の加速が期待される。

追加学習モデル名	ベース基盤モデル	ライセンス
Weblab-MedLLM-GLM-4.7	GLM-4.7	MIT
Weblab-MedLLM-gpt-oss-120b	gpt-oss-120b	Apache-2.0
Weblab-MedLLM-Qwen3-235B-Instruct	Qwen3-235B（Instruct 系）	Apache-2.0
Weblab-MedLLM-Qwen3-235B-Thinking	Qwen3-235B（Thinking 系）	Apache-2.0

追加学習モデルとライセンス

概要

東京大学大学院工学系研究科技術経営戦略学専攻／附属人工物工学研究センター 松尾・岩澤研究室（以下、松尾研）は、2026 年 5 月 28 日に公表した研究開発成果「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―」（※1）における主要な成果物のうち、松尾研が中心となって開発した医療特化型 LLM 群および安全性検証ツールを公開することをお知らせします。

5 月 28 日のプレスリリース（※1）では、NEDO「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」（※2）の成果として、患者情報を安全に管理できる環境で運用可能な医療業務支援 LLM が、専門医試験を模した学術試験において最大 90.8%（RAG 使用時）の正答率を達成し、世界最先端の商用 LLM（91.4%）に迫る性能と、日本の医療特性を踏まえた安全性の両立を実現したことをご報告しました。

今回は、その成果の中核を担う追加学習モデル群、国産フルスクラッチ開発モデル「AscleLM-1-10B」、ならびに安全性検証で実際に用いた 2 つのソフトウェアを、医療 LLM の社会実装の加速を目的に公開いたします。

発表内容

医療特化型 LLM（追加学習モデル）：

公開されているオープン基盤モデルに対し、日本の診療ガイドライン・専門医試験問題・臨床事例など、医療分野の教材から生成したデータを用いて追加学習（※3）を行ったモデル群です。商用利用可能な Apache-2.0、MIT ライセンスで以下のモデルを Hugging Face 上で公開します。

表 1：追加学習モデルとライセンス

追加学習モデル名	ベース基盤モデル	ライセンス
Weblab-MedLLM-GLM-4.7	GLM-4.7	MIT
Weblab-MedLLM-gpt-oss-120b	gpt-oss-120b	Apache-2.0
Weblab-MedLLM-Qwen3-235B-Instruct	Qwen3-235B（Instruct 系）	Apache-2.0
Weblab-MedLLM-Qwen3-235B-Thinking	Qwen3-235B（Thinking 系）	Apache-2.0

各モデルの概要、評価結果、利用条件については、Hugging Face 上の各モデルカードをご確認ください。

公開ページ：
Weblab-MedLLM-gpt-oss-120b：
https://huggingface.co/weblab-LLM-M/Weblab-MedLLM-gpt-oss-120b
Weblab-MedLLM-GLM-4.7：
https://huggingface.co/weblab-LLM-M/Weblab-MedLLM-GLM-4.7
Weblab-MedLLM-Qwen3-235B-Instruct：
https://huggingface.co/weblab-LLM-M/Weblab-MedLLM-Qwen3-235B-Instruct
Weblab-MedLLM-Qwen3-235B-Thinking：
https://huggingface.co/weblab-LLM-M/Weblab-MedLLM-Qwen3-235B-Thinking

国産フルスクラッチ開発モデル「AscleLM-1-10B」：

既存モデルを基にせず、設計から学習までを一から行うフルスクラッチ開発（※4）により構築した、独自アーキテクチャによる国産医療 LLM です。5 月 28 日プレスリリース（※1）に記載のとおり、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ており、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た成果物として、研究目的で公開します。

公開ページ：https://huggingface.co/weblab-LLM-M/AscleLM-1-10B

安全性検証ツール：

5 月 28 日プレスリリース（※1）に記載した、日本の医療特性を踏まえた多面的な安全性検証のうち、「レッドチーミングによる攻撃耐性評価」および「患者情報の自動検出・マスキング

機能」に用いたソフトウェア群を、研究者・開発者が自身のモデル・サービスにおいて同等の検証および安全運用を実施できるよう、公開します。

（1）レッドチーミング・アプリケーションおよびレポート

5 月 28 日プレスリリース（※1、図 2）に示したレッドチーミング（※5）に用いたアプリケーション本体と、評価結果を整理した集計レポートを提供します。医療領域特有のリスクシナリオに沿った敵対的プロンプトの設計・実行・結果分類・集計までを統合的に支援し、LLM 開発者が自身のモデルに対する同等の評価を再現・拡張できるよう設計されています。

公開リポジトリ：https://github.com/weblab-llm-m/red-teaming

（2）PII（個人情報）検出・ルーティングプログラム

5 月 28 日プレスリリース（※1）に記載した「患者情報を自動で検出・マスキングする機能」として実装されたプログラムを提供します。医療 LLM への入力プロンプトに含まれる氏名・連絡先等の Personally Identifiable Information（PII）を検出し、用途・リスクに応じてマスキング処理や別系統への振り分け等のルーティングを行うものです。医療 LLM の前段に組み込むことで、入力データの最小化原則に基づく安全な運用を支援することを目的としています。

公開リポジトリ：https://github.com/weblab-llm-m/pii-router

今後の展望：

今回のモデル本体と安全性検証ツールの公開により、LLM の研究者・開発者による再現・検証・発展を後押しし、医療業務支援 LLM の安全な社会実装の加速、医療現場の DX・医療の質向上につながることが期待されます。

本モデルの利用について

本プレスリリースで公開する追加学習モデル、フルスクラッチ開発モデル、および安全性検証ツールは、医療従事者の事務作業・文書作成等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うことを前提とします。各リポジトリおよびモデルカードをご確認の上、ご利用ください。

松尾・岩澤研究室について

東京大学 松尾・岩澤研究室では、「知能を創る」ことをビジョンに掲げ、ディープラーニングの研究を推進しています。特に、世界モデルやロボット研究、大規模言語モデル、脳×AI に関する研究を進めています。加えて、基礎研究成果を社会に還元することにも注力しており、講義、企業との共同研究、学生起業家の育成支援なども行っています。

注釈

（※1）「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―」（2026 年 5 月 28 日プレスリリース）
https://www.t.u-tokyo.ac.jp/press/pr2026-05-28-001

（※2）NEDO（新エネルギー・産業技術総合開発機構）事業「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」（事業期間：2025 年度）
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。

（※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。

（※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、システムの脆弱性を体系的に評価する手法。

※本取り組みは、NEDO 事業（※2）の支援を受けて実施された研究開発成果に基づくものです。

問合せ先

東京大学 大学院工学系研究科
技術経営戦略学専攻 松尾・岩澤研究室
広報担当