NEWS RELEASE

大学共同利用機関法人 情報・システム研究機構 国立情報学研究所 National Institute of Informatics

2026 年（令和 8 年）10 月 26 日

国産医療 LLM と 3 万例規模の疾患因果ダイアグラムを統合した医療 AI エージェントを開発 ～医師国家試験ベンチマークで 90%超を達成した「SIP-jmed-llm-4」を活用、医師向け試験公開へ～

医療現場では、患者の症状や検査所見から複数の疾患を検討する必要があり、膨大な医学知識や過去の症例情報を効率的に参照できる診断支援技術が求められています。近年、こうした課題に対し、大規模言語モデル（LLM）が外部の情報やツールを活用して診断支援を行う「医療 AI エージェント」が注目されています。

これまで、学校法人自治医科大学の永井 良三学長（内閣府・戦略的イノベーション創造プログラム（SIP）(*1) 第 3 期研究開発課題「統合型ヘルスケアシステムの構築」(*2) プログラムディレクター）は、約 3 万例の症例について疾患・病態と症状・所見の関係を整理した「疾患因果ダイアグラム」(*3) と医学用語辞書を長年にわたり構築してきました。

今回、大学共同利用機関法人情報・システム研究機構 国立情報学研究所（NII、所長：黒橋 禎夫、東京都千代田区）を中心とする研究チームは、SIP「統合型ヘルスケアシステムの構築」の支援のもと、国産医療 LLM「SIP-jmed-llm-4」を開発し、医師国家試験ベンチマーク「JMedQA」において 90%を超える正答率を達成しました。さらに、疾患因果ダイアグラムに基づく症例知識グラフを外部ツールとして参照する医療 AI エージェントを開発し、知識グラフを参照しない場合と比べて、鑑別診断（症状や検査所見から考えられる複数の疾患を挙げ、比較して絞り込むこと）の性能が向上することを確認しました。

NII は、これらの技術を統合し、関連する症例情報の検索・参照機能を備えた臨床意思決定支援 Web アプリケーションを開発しました。医師による効率的な情報収集や鑑別診断の支援への貢献が期待されます。今後、医師による試験利用を通じて有用性や安全性を検証するとともに、関係機関との連携のもと、権利関係や利用条件が整理された範囲で 2026 年度内の試験的な限定公開を目指します。

[背景・課題]

大規模言語モデル（LLM）の医療応用が急速に進展し、主に海外で開発された LLM が、医師国家試験などの医学知識を問うベンチマークにおいて高い性能を示しています。一方、国民の健康と生命に直結する医療分野では、日本の言語・医療制度・診療実践に根ざした高性能な LLM を国内で主体的に開発し、その技術や開発過程を把握・管理できる体制を確立することが、医療 AI の技術的主権を確保するうえで重要な課題となっています。

National Institute of Informatics

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
総務部総務企画課 企画・広報チーム
〒101-8430 東京都千代田区一ツ橋 2-1-2
E-Mail: media@nii.ac.jp

Web: https://www.nii.ac.jp
X: @jouhouken
facebook: https://www.facebook.com/jouhouken

さらに、実際の診療では、幅広い医学知識に加え、患者の症状、検査所見、病歴などを総合的に評価し、過去の症例に基づく情報を参照しながら、複数の疾患候補を検討することが求められます。

このため、①日本の言語・医療制度・診療実践に適した高性能な国産医療 LLM、②実際の症例に基づく医学的知識を外部ツールとして参照・活用する仕組み、③これらを統合し、医師が利用・評価できる医療 AI エージェントシステムという 3 つを一体として整備することが重要です。

これまで、自治医科大学の永井学長は、長年にわたり、約 3 万例の症例について疾患・病態と症状・所見の関係を整理・構造化した「疾患因果ダイアグラム」と、医学用語を体系的に整理した医学用語辞書を構築してきました。また、株式会社プレシジョン（代表取締役・医師：佐藤 寿彦）との共同研究を通じて、これらの知識基盤を活用するためのシステム開発や用語辞書の管理・運用が進められてきました。これらは、実際の症例に基づく医学的知識や医学用語を体系的に整理した知識基盤であり、医療 AI が参照・活用できる医学知識資源としての応用が期待されています。

これと並行して、NII の大規模言語モデル研究開発センター（LLMC）では、日本語に強く、開発過程の透明性を重視した国産基盤モデル「LLM-jp」シリーズ (*4) を開発しています。この LLM-jp を基盤として、NII コンテンツ科学研究系の相澤 彰子研究室を中心とする研究チームは、上記 SIP 課題の支援のもと、ソブリン医療 LLM (*5)「SIP-jmed-llm」の研究開発を進めてきました。

また、上記 SIP 課題のテーマ E-2「大規模医療文書・画像の高精度解析基盤技術の開発」（研究開発責任者：NII 黒橋 禎夫）では、NII および京都大学大学院情報学研究科を中心とする研究チームが、疾患因果ダイアグラムを活用し、症例報告文書から疾患、症状、検査所見、病態生理学的な関係などを自動的に抽出・構造化する技術の研究開発を進めてきました。

今回、上記 SIP 課題のテーマ B-6「大規模症例知識ベースと統合した実践型医療 LLM の開発と社会実装」（研究開発責任者：NII 小林 和馬特任准教授）では、これらの研究成果を発展させ、**国産医療 LLM「SIP-jmed-llm-4」を開発**するとともに、**疾患因果ダイアグラムに基づく大規模症例知識グラフを推論時に外部ツールとして参照・活用し、症例情報に応じた鑑別診断を支援する医療 AI エージェントを構築**しました。さらに、これに症例情報の検索・参照機能を組み合わせ、**臨床意思決定支援 Web アプリケーションとして実装**し、医師による試験利用に向けた技術的基盤を整備しました。

[研究・開発／取り組みの概要]

NII を中心とする研究チームは、国産医療 LLM「SIP-jmed-llm-4」を開発するとともに、疾患因果ダイアグラムに基づく大規模症例知識グラフを外部ツールとして参照・活用する医療 AI エージェントを構築しました。この AI エージェントは、患者の症状や検査所見に応じて関連する医学知識を取得し、鑑別診断を支援します。さらに、関連する症例報告文書を検索・参照できる機能を組み合わせ、主に医師による試験利用を想定した臨床意思決定支援 Web アプリケーションとして実装しました（図 1）。

SIP-jmed-llm-4 は、NII LLMC が開発した国産基盤モデル「LLM-jp-4-33b-base」をベースに、日本語の医療分野に特化した継続事前学習および事後学習を施して構築されました。約 332 億パラメータからなる医療特化 LLM であり、外部ツールを利用して情報を取得・活用するエージェント機能を備えています。日本の医師国家試験問題からなるベンチマーク「JMedQA」(*6) では、外部ツールを用いない条件で 90%を超える正答率を達成しました（図 2）。

開発した医療 AI エージェントは、入力された症例情報に応じて、SIP-jmed-llm-4 が疾患因果ダイアグラムに基づく症例知識グラフを推論時に外部ツールとして参照し、疾患と症状・検査所見の関係や共起情報など、症例から抽出・構造化された医学的知識を取得します。また、Web アプリケーションには、利用条件を満たす症例報告データベースと連携し、関連する症例報告文書を検索・参照する機能も実装

しています。これにより、LLM 内部に学習された医学知識に加えて、外部の構造化知識や関連する症例報告文書を活用し、複数の疾患候補の検討を支援する一体型の臨床意思決定支援システムを実現しました。

また、医療 AI エージェントの鑑別診断性能を評価するため、公開された症例報告データをもとに、日本語鑑別診断ベンチマーク「PMC-Patients-JA-DDx」(*7) を構築しました。本ベンチマークを用いて、SIP-jmed-llm-4 が疾患因果ダイアグラムに基づく症例知識グラフを外部ツールとして利用する場合と、利用しない場合の鑑別診断性能を比較・検証しました（表 1）。

[主な特徴・新規性／技術的内容]

本成果の特徴は、高性能な国産医療 LLM の開発に加え、約 3 万例の症例から整理された医学知識を外部ツールとして活用する医療 AI エージェントを構築し、医師が試用できる臨床意思決定支援 Web アプリケーションとして実装した点にあります。主なポイントは以下の 3 点です。

- **高性能な国産医療 LLM「SIP-jmed-llm-4」を開発**
  国産基盤モデル「LLM-jp-4」をベースに、日本語医療コーパスによる継続事前学習と事後学習を行い、外部ツールを利用できる医療特化 LLM を開発しました。医師国家試験ベンチマーク「JMedQA」では、外部ツールを用いない条件で 90%を超える正答率を達成しました。

- **約 3 万例の症例に基づく疾患因果ダイアグラムを活用した医療 AI エージェントを構築**
  疾患因果ダイアグラムに基づく大規模症例知識グラフを外部ツールとして参照し、疾患と症状・検査所見の関係や共起情報などを取得する医療 AI エージェントを構築しました。これにより、LLM 内部の医学知識に加え、実際の症例に基づく構造化知識を活用した鑑別診断支援を実現し、ベンチマーク評価において鑑別診断性能の向上を確認しました。

- **症例報告データベースと柔軟に連携できる Web アプリケーションを実装**
  医療 AI エージェントに加え、関連する症例報告文書を検索・参照できる機能を備えた臨床意思決定支援 Web アプリケーションを開発しました。特定のデータベースに依存せず、利用条件を満たす症例報告データベースと柔軟に連携できる構成とし、医師による試験利用に向けた技術的基盤を整備しました。

[得られた成果・効果]

SIP-jmed-llm-4 は、外部ツールを用いない条件で、医師国家試験ベンチマーク「JMedQA」において 90%を超える正答率を達成しました。また、モデルの規模を示す総パラメータ数と正答率を比較したところ、今回評価した日本語特化モデル群の中で、モデル規模と性能のバランスに優れた「パレートフロンティア」上に位置しました（図 2）。パレートフロンティアとは、モデル規模の小ささと正答率の高さの両面で、他のモデルに同時に上回られないモデル群が形成する境界を指します。この結果は、SIP-jmed-llm-4 がモデル規模に対して医療分野における高い問題解答能力を備えており、国内で開発過程を管理できるソブリン医療 LLM として、実用的な医療 AI システムの中核となり得る基礎性能を有することを示しています。

また、医療 LLM の医学的正確性や安全性・信頼性を高めるためのアライメント技術についても研究開発を進めています。その成果の一つである、モデルの回答をより適切なものへと改善する選好学習手法に関する研究が、自然言語処理分野の主要国際会議 EMNLP 2026 の Findings に採択されました(*8)。

さらに、評価用に構築した日本語鑑別診断ベンチマーク「PMC-Patients-JA-DDx」を用いて医療 AI エージェントの性能を評価したところ、疾患因果ダイアグラムに基づく症例知識グラフを外部ツールとして利用した SIP-jmed-llm-4 は、同じモデルで知識グラフを利用しない場合よりも高い鑑別診断性能を示しました（表 1）。この結果は、実際の症例から抽出・構造化された医学的知識を推論時に参照・活用することが、医療 AI エージェントによる鑑別診断支援の性能向上に有効である可能性を示しています。

なお、これらの結果は研究用ベンチマークに基づく評価であり、そのまま実際の診療における安全性や有用性を保証するものではありません。今後、医師による試験利用と専門的な評価を通じて、医療 AI エージェントの有用性や安全性を継続的に検証していく必要があります。

[想定される活用分野・利用シーン]

主に医師による利用を想定し、患者の症状や検査所見から複数の疾患の可能性を検討する場面での活用を目指します。SIP-jmed-llm-4 が内部に保持する医学知識に加え、疾患因果ダイアグラムに基づく症例知識グラフや、連携する症例報告データベースから検索された関連文書を参照・活用することで、医師が鑑別疾患の候補や関連する医学的情報を効率的に確認できるよう支援します。また、医師による試験利用を通じて、外部の医学知識を活用する医療 AI エージェントの有用性や安全性を検証し、専門家からのフィードバックを継続的な研究開発に還元するための研究基盤としても活用します。

本システムは研究開発段階のプロトタイプであり、医師の診断や判断を代替することを目的としたものではありません。

[社会的意義・期待される影響]

日本の医療には、日本語だけでなく、国内の医療制度や診療の実践に根ざした知識が求められます。本研究では、国産基盤モデル「LLM-jp」を基盤として日本の医療に特化した SIP-jmed-llm-4 を開発し、さらに疾患因果ダイアグラムに基づく症例知識グラフと外部の症例報告データベースを活用する医療 AI エージェントを構築しました。これにより、国内で主体的に開発・管理し、継続的に高度化できる医療 AI 基盤の実現に貢献することが期待されます。

また、外部の医学知識や症例報告文書を参照しながら回答することで、複数の疾患候補の検討を支援するとともに、回答の根拠となる情報を医師が確認しやすい医療 AI の実現を目指します。

[今後の展開・予定]

今後、2026 年度内を目標に、SIP-jmed-llm-4 と疾患因果ダイアグラムに基づく症例知識グラフを統合し、利用条件を満たす症例報告データベースと連携可能な臨床意思決定支援 Web アプリケーションを、主に医師を対象として試験的に限定公開する予定です。医師による試験利用を通じて有用性や安全性を検証するとともに、得られたフィードバックをもとに、医療 LLM の性能、知識グラフの活用方法、症例検索、回答生成、ユーザーインターフェースなどを継続的に改善し、医療 AI エージェントの高度化を進めます。Web アプリケーションの詳細および試験公開に関する最新情報は、上記 SIP 課題のテーマ B-6 の Web サイト（https://sites.google.com/nii.ac.jp/sip3b-6/）で公開します。

さらに、SIP-jmed-llm-4 の継続的な高度化に加え、疾患因果ダイアグラムに基づく症例知識グラフの拡充や、診療ガイドライン、医学文献などの信頼性の高い医学情報を必要に応じて参照・活用する機能の強化を進めます。また、上記 SIP 課題のテーマ E-2 とテーマ B-6 の連携のもと、匿名加工医療情報を活用した仮想電子カルテ環境を構築し、診療経過に伴って蓄積される時系列情報を利用する医療 AI エージェントの研究開発へと展開します。これらの取り組みにあたっては、必要な倫理審査や関連法令・

指針を遵守し、医師による実用性、安全性、医学的妥当性の評価を段階的に進め、将来の臨床応用と社会実装につなげることを目指します。

図 1　臨床意思決定支援 Web アプリケーションの開発画面

左側には症例報告データベースから検索された関連症例の一覧と詳細情報、中央には疾患や症状・検査所見の関係を探索するための知識グラフ、右側には入力症例に関する質問と、SIP-jmed-llm-4 を用いた医療 AI エージェントの回答を表示する。

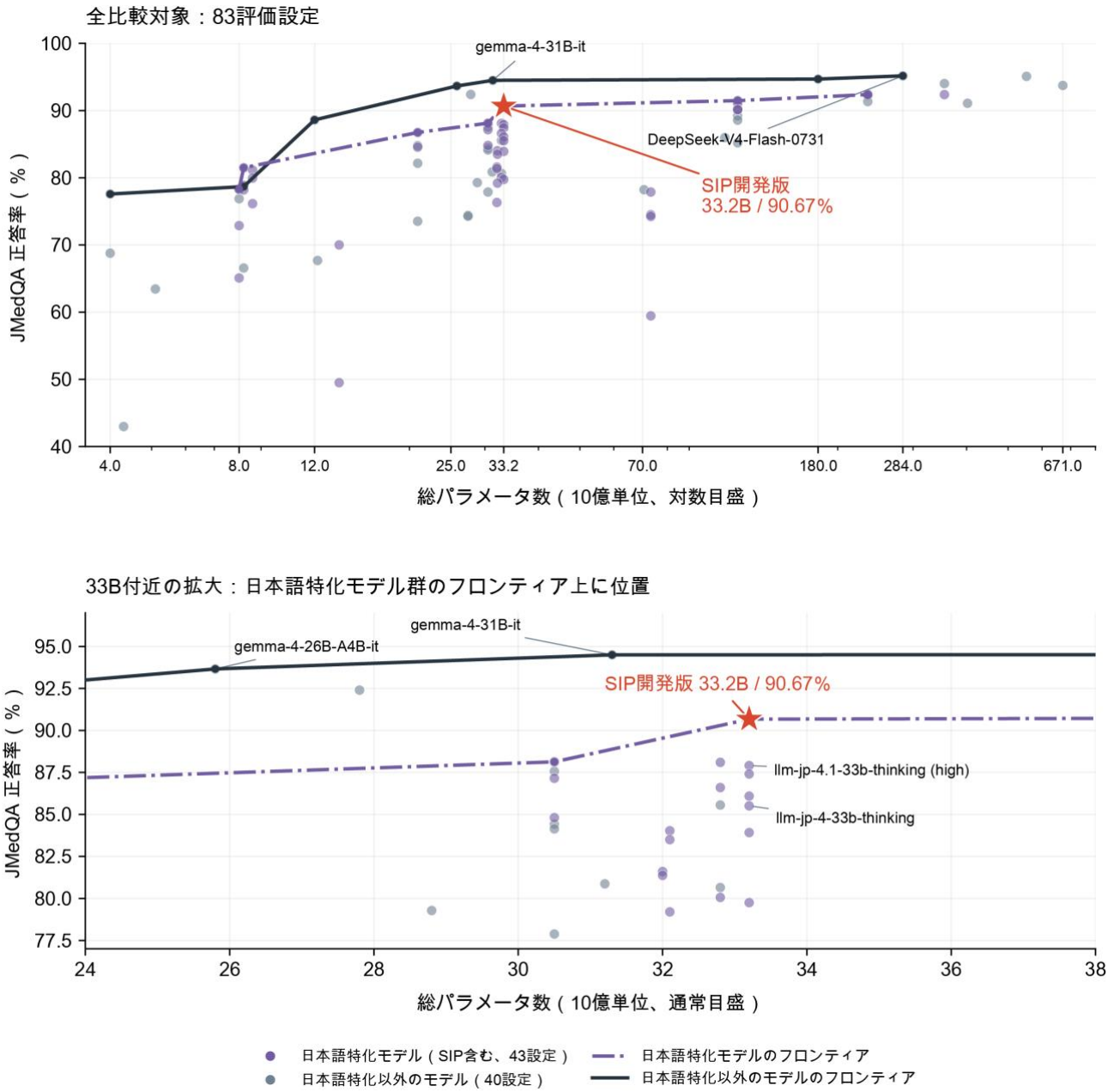

図 2　JMedQA における総パラメータ数と正答率の比較

赤い星は SIP-jmed-llm-4（SIP 開発版）を示す。上段は全比較モデル、下段は約 330 億パラメータ付近の拡大図。各曲線は、それぞれのモデル群におけるパレートフロンティアを示す。
SIP-jmed-llm-4 は今回比較した日本語特化モデルのうち、総パラメータ数 332 億以下では最高の正答率を達成し（紫色の点線）、日本語特化モデル群のパレートフロンティア上に位置した。

モデルと評価構成	平均評価点（3 点満点）
SIP-jmed-llm-4（症例知識グラフを外部ツールとして利用）	1.97
gpt-oss-120b	1.93
SIP-jmed-llm-4（症例知識グラフを利用しない）	1.74
gpt-oss-20b	1.67
llm-jp-4-33b-thinking	1.58
Qwen3-32B	1.57
Qwen3-30B-A3B-Thinking	1.47
Qwen3-30B-A3B-Instruct	1.40

表 1　PMC-Patients-JA-DDx の平均評価点（3 点満点）

SIP-jmed-llm-4 について、疾患因果ダイアグラムに基づく症例知識グラフを外部ツールとして利用する場合と、利用しない場合の鑑別診断性能を比較。知識グラフを利用した構成では、利用しない構成よりも高い平均評価点を示した。

謝辞

本成果は、内閣府・戦略的イノベーション創造プログラム（SIP）第 3 期研究開発課題「統合型ヘルスケアシステムの構築」（プログラムディレクター：学校法人自治医科大学　永井　良三　学長、課題番号：JPJ012425）の支援を受け、その研究開発テーマ E-2 およびテーマ B-6 における研究開発を通じて得られたものです。

上記 SIP 課題のテーマ E-2 における SIP-jmed-llm の研究開発には、東京科学大学情報理工学院　荒瀬　由紀教授、京都大学大学院情報学研究科　村脇　有吾准教授、程　飛特定講師、八幡　早紀子博士課程学生をはじめ、多くの研究者にご協力いただきました。

また、SIP-jmed-llm-4 の開発から、疾患因果ダイアグラムに基づく症例知識グラフとの統合、性能評価、臨床意思決定支援 Web アプリケーションの実装に至る一連の研究開発には、NII 劉　超然特任准教授、小田　悠介特任准教授、清丸　寛一特任助教、児玉　貴志特任助教、江　俊鋒特任研究員、Wu Yun-Ang　特任研究員、柴木　亮太外来研究員、坂井　梓外来研究員、山岸　陽助リサーチ・アシスタント、渡辺　悠介リサーチ・アシスタント、目時　大暉リサーチ・アシスタントをはじめ、多くの研究者・研究支援者が参画しました。

〈メディアの皆様からのお問い合わせ先〉
大学共同利用機関法人　情報・システム研究機構
国立情報学研究所
総務部総務企画課　企画・広報チーム
E-mail：media@nii.ac.jp

[用語解説]

（*1）戦略的イノベーション創造プログラム（SIP）

戦略的イノベーション創造プログラム（SIP）とは、政府の総合科学技術・イノベーション会議（CSTI）が司令塔機能を発揮して、府省の枠や旧来の分野を超えたマネジメントにより、科学技術イノベーション実現のために創設した国家プロジェクトです。真に重要な社会的課題や、日本経済再生に寄与できるような世界を先導する課題に取り組んでいます。各課題を強力にリードするプログラムディレクター（PD）を中心に産学官連携を図り、基礎研究から実用化・事業化、すなわち社会実装までを見据えて一気通貫で研究開発を推進しています。

内閣府「戦略的イノベーション創造プログラム（SIP）」公式サイト：
https://www8.cao.go.jp/cstp/gaiyo/sip/

（*2）統合型ヘルスケアシステムの構築

統合型ヘルスケアシステムの構築とは、戦略的イノベーション創造プログラム（SIP）第 3 期の研究開発課題の一つであり、医療デジタルツインの実装により医療・ヘルスケアにおける「知識発見」と「医療提供」の循環が自律的に促進され、医療の質向上、健康寿命延伸、医療産業振興、持続可能な医療制度に活用されることを目指しています。そのために、医療デジタルツインに必要な「技術開発」と医療デジタルツインを活用した医療・ヘルスケア課題のソリューションの「事業」化に取り組みます。

※デジタルツインとは、現実世界（物理空間）の情報をデジタル化し、仮想空間（デジタル空間）上に再現したモデルを構築し、現実に近いシミュレーションを行い、社会にフィードバックする仕組みのことです。

SIP 第 3 期研究開発課題「統合型ヘルスケアシステムの構築」公式サイト：
https://sip3.jihs.go.jp/index.html

（*3）疾患因果ダイアグラム

自治医科大学の永井 良三学長が、長年にわたり約 3 万例の症例について、疾患・病態と症状・検査所見の関係を整理・構造化した医学知識基盤。約 5 万語の医学用語辞書とともに構築され、症例に含まれる医学的な関係性を体系的に整理したものです。

また、自治医科大学と株式会社プレシジョンとの共同研究を通じて、これらの知識基盤を活用するための症例検索システムの開発や、用語辞書の管理・運用が進められてきました（平成 28 年度 AMED 臨床研究等 ICT 基盤構築研究事業「人工知能による総合診療診断支援システムの開発」（研究代表者：永井 良三）、および令和元年度 NEDO Connected Industries 推進のための協調領域データ共有・AI システム開発促進事業「医療情報を横断的に統合した診療支援 AI システムの開発」（研究代表者：佐藤 寿彦））。

（*4）LLM-jp シリーズ

NII LLMC を中心に開発している、日本語に強い国産大規模言語モデルのシリーズ。本研究では、最新世代の「LLM-jp-4」を SIP-jmed-llm-4 の基盤モデルとして利用しています。

（*5）ソブリン医療 LLM

基盤モデルの開発、学習、評価、運用に関する技術や過程を国内主体で把握・管理し、継続的に改良可能な医療分野の大規模言語モデル。

（*6）JMedQA

日本の医師国家試験問題をもとに、大規模言語モデル等の医学知識・問題解決能力を評価するために構築されたベンチマーク。
https://huggingface.co/datasets/SIP-med-LLM/JMedQA

（*7）PMC-Patients-JA-DDx

PubMed Central（PMC）の症例報告等から構築された「PMC-Patients」をもとに、症例情報から鑑別診断を行う能力を日本語で評価するために構築したベンチマーク。
https://huggingface.co/datasets/SIP-med-LLM/PMC-Patients-JA-DDx

[関連研究]

（*8）Kazuma Kobayashi, Yosuke Yamagishi, Ryota Shibaki, Azusa Sakai, Takashi Kodama, Lin Gu, Irene Li, Yuki Arase, Akiko Aizawa, and Sadao Kurohashi. “Stabilizing Hard-Negative Preference Optimization for Medical LLMs.” Findings of the Association for Computational Linguistics: EMNLP 2026, 2026.