

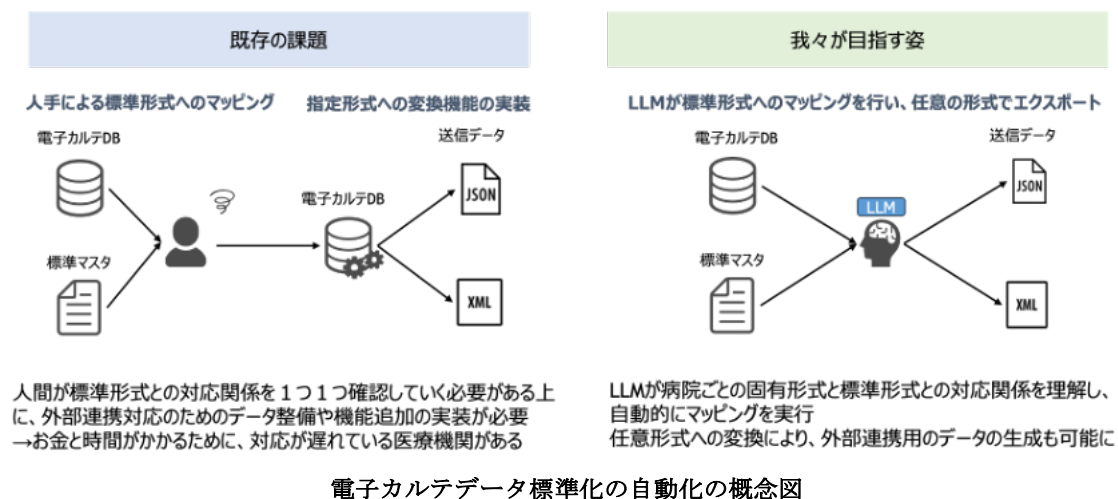
※注釈のうち(※2)を修正しました。(2026年3月6日)

東京大学大学院工学系研究科

東京大学 松尾・岩澤研究室、医療現場の DX の実現を目指し 日本語版医療特化型 LLM を開発し、対話型 AI サービスを公開

発表のポイント

- ◆ 日本語医学知識を付与した医療特化型 LLM を開発し、2025 年医師国家試験ベンチマークで正答率 93.3%を達成。
- ◆ 電子カルテデータ標準化を想定したユースケース検証では、標準名称への変換タスクにおいて国内外の LLM の性能を上回った。
- ◆ 本成果により、医療データ標準化や業務効率化の自動化が進み、医療現場の DX 推進や創薬・治験の高度化を通じて、日本の医療・創薬力向上への貢献が期待される。



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）が、さくらインターネット株式会社、株式会社 ELYZA、株式会社 ABEJA、理化学研究所および医療機関と連携して開発した日本語版医療特化型 LLM を研究目的限定で研究者向けに提供することをお知らせします。

本モデルは、2025 年の医師国家試験ベンチマークにて OpenAI 社の OpenAI-o1 や GPT-4o を上回る 93.3%の正答率を記録しました。さらには、臨床現場における具体的なユースケースを想定し、LLM による電子カルテデータ標準化の自動化の検証も実施しています。本研究は、戦略的イノベーション創造プログラム (SIP) 統合型ヘルスケアシステムの構築 (JPJ012425) の補助を受けて実施されたものです(※1)。

これらの成果は、国内の AI 技術の高さを証明するとともに、人手不足や業務の複雑化といった課題を抱える医療現場のデジタルトランスフォーメーション (DX) を大きく前進させる可能性を秘めています。以下では、本取り組み内容の詳細について紹介します。

発表内容

日本語版医療特化型 LLM の開発 :

松尾研は、オープン LLM である Qwen-2.5-72B-Instruct をベースモデルとし、upcycling によるモデルサイズ拡張や医学論文等の医学系コーパスを用いた継続事前学習と指示学習を重ねることで、日本語医学知識を付与した Weblab-MedLLM-Qwen-2.5-109B-Instruct を構築しました。図 1 で示すように、本モデルは既存モデルが保持していない日本国内の医療制度に関する知識を備えており、既存モデルが誤答してしまう問題にも正しく答えることが可能です。表 1 のとおり、本モデルは 2025 年医師国家試験ベンチマークで正答率 93.3% を記録しており、これは同ベンチマークにおいて OpenAI 社の GPT-4o や OpenAI-o1 を上回るトップクラスの性能を示しています。

さらに、外部の知識データベースを参照する RAG (Retrieval-Augmented Generation) や多数決によって回答精度を高める majority voting などの技術を組み合わせることで、正答率は最大で約 98% にまで向上することも確認されました (ただし、図の参照を必要とする問題や計算問題などを除く)。

これらの結果から、医学系コーパスを用いた追加学習を実施することで、オープン LLM に日本語の医学知識を付与し、医師国家試験などの高度な専門知識を要する問題に対しても極めて高い精度で回答できることが明らかになりました。

設問例 (正解はd)

介護保険による機能訓練で正しいのはどれか。

a: 介護福祉士が実施する。b: 利用者は減少している。c: 医師の指示が必要である。d: 家事動作訓練が含まれる。e: 特定機能病院で実施される。a, b, c, d, e の中から1つ選びなさい。

OpenAI-o1の回答

介護保険による機能訓練は、リハビリテーション専門職 (理学療法士・作業療法士・言語聴覚士など) または機能訓練指導員の資格を持つ者が、医師の指示に基づいて実施されます。したがって、

- a「介護福祉士が実施する」は誤り (介護福祉士が主導して行うものではない)
- b「利用者は減少している」も誤り (要介護高齢者数はむしろ増加傾向)
- c「医師の指示が必要である」は正しい
- d「家事動作訓練が含まれる」は、実際には日常生活活動 (ADL) や IADL を向上させるために調理や掃除などの訓練を行うことはありますが、選択肢の中で「最も確実に正しい」とされるのは医師の関与を明記している c です
- e「特定機能病院で実施される」は誤り (特定機能病院は高度医療を提供する病院であり、介護保険下の機能訓練として一般的に行われる場ではない)

以上より、正しいのは「医師の指示が必要である」すなわち c です。

Weblab-MedLLM-Qwen-2.5-109B-Instructの回答

問題文は「介護保険による機能訓練で正しいのはどれか」であり、選択肢 a~e の中から一つを選ぶものです。各選択肢について検討していきます。

まず、**a: 介護福祉士が実施する**ですが、介護保険制度における機能訓練の実施者は、理学療法士 (PT)、作業療法士 (OT)、言語聴覚士 (ST) などリハビリ専門職が主となります。介護福祉士は直接的な機能訓練を行う資格がないため、この選択肢は誤りです。

次に**b: 利用者は減少している**ですが、日本の高齢化率は上昇傾向にあり、介護サービス需要も増加しています。そのため、機能訓練を含む介護保険サービスの利用者が減少しているとは考えにくく、こちらも誤りでしょ。

c: 医師の指示が必要であるについては、介護保険下でのリハビリテーションは医療保険とは異なり、必ずしも医師の指示が必須ではないとされています。医療保険でのリハビリでは医師の指示が必要ですが、介護保険の場合はケアプランに基づいて実施され、直接的に医師の指示が毎回必要というわけではありません。よって c は誤りです。

d: 家事動作訓練が含まれるに関しては、介護保険の機能訓練は日常生活動作 (ADL) や手段的日常生活動作 (IADL) の改善を目指すため、家事動作訓練も対象となります。例えば、作業療法士が調理や掃除などの動作訓練を行いますので、d は正しいと考えられます。

最後に**e: 特定機能病院で実施される**ですが、特定機能病院は高度医療を提供する施設であり、介護保険の機能訓練は在宅や介護施設、デイサービスなどで行われることが一般的です。特定機能病院での実施は通常ではなく、e は誤りです。

以上より、正解は **d** となります。

図 1 : 医師国家試験の設問と解答の具体例

表 1 : 2025 年医師国家試験の正答率比較表

モデル	分類	事業者	正答率
Weblab-MedLLM-Qwen-2.5-109B-Instruct	国内	Ours	93.3%
OpenAI-o1	国外	OpenAI	92.8%
DeepSeek R1	国外	DeepSeek	91.5%

Preferred-MedLLM-Qwen-72B	国内	Preferred Networks	82.0%
GPT-4o	国外	OpenAI	88.5%

ユースケース検証：

実際の医療現場を想定した電子カルテデータ標準化のユースケース（図 2）を想定し、感染症・検査情報の名称を厚労省の定める標準名称に変換するタスクにも取り組みました。その結果、本モデルは図 3 のように既存モデルでは回答が難しい問題に対しても、獲得した医学知識に基づいて正しい標準名称を導出可能であることが示されました。表 2 のとおり、モデルは F1 スコア 85% の精度で標準名称への変換が可能であり、標準化のタスクにおいても国内外の LLM の性能を大きく上回っています。

これらの結果から、医学知識を付与した LLM によって非常に高い精度で電子カルテデータ標準化が実現可能であり、LLM による標準形式への自動マッピングによって医療現場の DX の実現に貢献できる可能性があることが示されました。

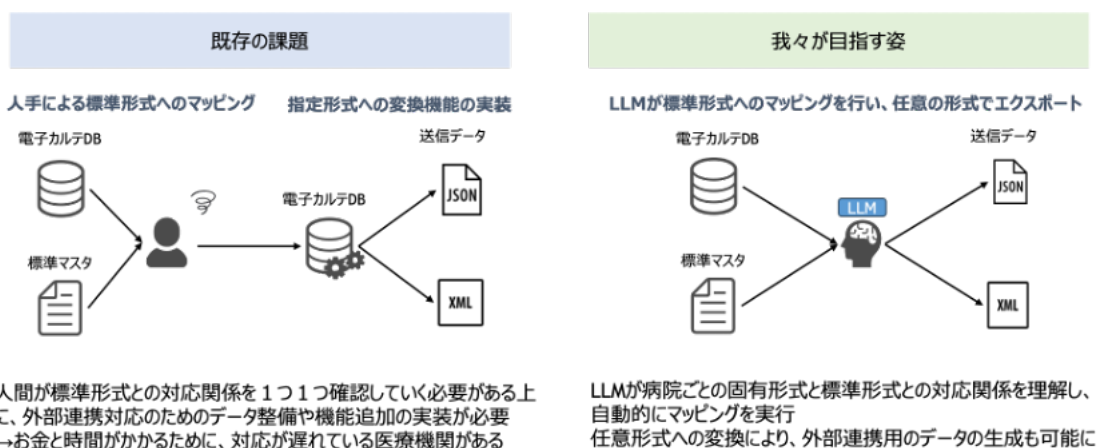


図 2：電子カルテデータ標準化の自動化の概念図

設問例（正解はアルカリホスファターゼ）

検査名: ALP 1 結果: %
 に対応する標準名称を以下のリストから選択せよ。

[標準名称リスト]

考察した後、最後に[ans]/ansタグで囲った回答を出力しなさい。

GPT-4oの回答

テスト項目「ALP 1」は、標準名称のリストに直接該当するものはありません。また、結果が「%」で示されているため、定量的な測定ではないと考えられます。したがって、標準名称として該当するものはないと判断します。

標準名称 [ans]該当なし[/ans]

Weblab-MedLLM-Qwen-2.5-109B-Instructの回答

<think>

まず、問題文を見てみましょう。「テスト項目: 検査名: ALP 1 結果: %」というデータに対して、標準名称を特定する必要があります。最初に、ALP (アルカリホスファターゼ) に関連する項目を考えます。ALPは通常、血液中の酵素活性を測定する検査で、その値が数値で表されます。しかし、今回の結果は「%」で示されています。これは少し不思議ですね。一般的にALPの単位はU/LやIU/Lが多いですが、「%」を使う場合、おそらく他の何かの比率や分画を示しているかもしれません。

...

通常は各アイソザイムの割合を計算するために、総ALPに対する比率を出すので、その場合は「アルカリホスファターゼ」の下位項目として扱われるかもしれません。しかし、標準名称リストにはそういった詳細な項目がないため、最も近い「アルカリホスファターゼ」を選択するのが妥当と思われる。

</think>

検査名「ALP 1」の結果が「%」で示されている点に注目する。アルカリホスファターゼ (ALP) のアイソザイム分画では、各アイソザイムの占める割合がパーセンテージで報告されることがある。標準名称リストにおいて「アルカリホスファターゼ」はALP全般を包括する項目であり、アイソザイム分画も含めて適用範囲に入る。類似事例 (例: 「ALP 2」→「アルカリホスファターゼ」) とも整合する。単位が「%」であっても、定量的な比率を表すため(定量)の区分には該当せず、基本項目として「アルカリホスファターゼ」が適切である。

[ans]アルカリホスファターゼ[/ans]

図 3：感染症・検査情報標準化の設問と解答の具体例

表 2：感染症・検査情報の標準名称への変換精度の比較表

モデル	分類	事業者	Accuracy	F1
Weblab-MedLLM-Qwen-2.5-109B-Instruct	国内	Ours	99%	85%
Qwen2.5-72B	国外	Alibaba	99%	78%
GPT-4o	国外	OpenAI	99%	77%
GPT-3.5-turbo	国外	OpenAI	89%	48%

今後の展望：

2025年度のNEDO事業（※2）においては、SIPでの取り組みを発展させ、複数の医療機関の電子カルテシステムと連携して動作するLLMエージェントとその安全性を評価する仕組みの構築を目指す予定です。これによって、労働集約的に行われている治験患者の探索やレジストリの構築を自動化し、製薬企業が必要とするデータを任意形式で提供することが可能となり、日本の創薬力の向上につながると期待されます。

本モデルの利用について

松尾研では研究成果の評価、研究のさらなる推進に貢献するため、開発したモデルを利用できる対話型AIサービスを提供します。なお、本サービスの提供については、予告なく変更または終了する場合がございますので、あらかじめご了承ください。

〈公開期間〉

2026年3月5日～2026年8月31日

〈対話型AIサービス上での利用〉

本モデルは、松尾研が提供する対話型AIサービス上で利用可能です。本サービスは、さくらインターネット株式会社の「さくらのAI Engine（※3）」上にデプロイした本モデルをチャットアプリから呼び出せるようにしたものです。

医学知識を学習させたモデルであるため、医学試験の問題や医学知識に関する質問を入力して確認していただくことが可能です。なお、本プログラムは診断行為、診療行為および治療行為に用いることはできません。

質問の例：

- ・「介護保険による機能訓練で正しいのはどれか。 a: 介護福祉士が実施する。 b: 利用者は減少している。 c: 医師の指示が必要である。 d: 家事動作訓練が含まれる。 e: 特定機能病院で実施される。 a,b,c,d,eの中から1つ選びなさい。考察した後、最後に[ans]/[/ans]タグで囲った回答を出力しなさい。ansタグ内には選択肢の文字のみを出力しなさい。」
- ・「HPVワクチンは安全ですか？」

本サービスの利用を希望される方は、利用規約をご確認の上でご利用ください。今後の研究プロジェクトの推進のため、本サービス上に入力されたプロンプト、プロンプトに対応するモデルの出力結果、モデルの出力結果に対応するフィードバックのログなどについては、本プログラムの学習に利用される可能性があることをご了承ください。

対話型サービス URL :

<https://weblab-medllm-qwen-25-109b-instruct.medllm.weblab.t.u-tokyo.ac.jp/>

(利用規約はアクセス時に表示されます)

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) 戦略的イノベーション創造プログラム (第3期) 「統合型ヘルスケアシステムの構築における生成 AI の活用」

<https://sip3.jih.go.jp/news/2024/publicrecruitmentresult.html>

(※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」

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

(※3) 生成 AI 向け推論 API 基盤 「さくらの AI Engine」

<https://www.sakura.ad.jp/aipf/ai-engine/>

※本取り組みは、戦略的イノベーション創造プログラム (SIP) 統合型ヘルスケアシステムの構築 (JPJ012425) の補助を受けて実施されました。

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 白石萌莉