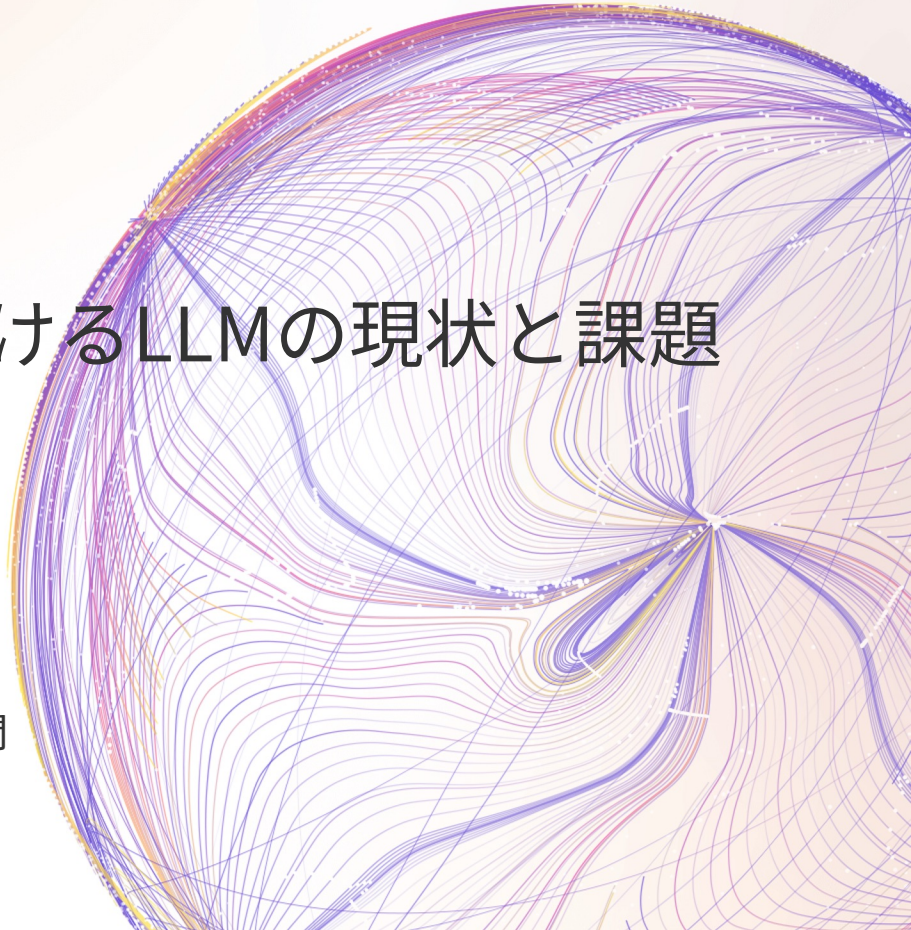
SB Intuitions

# 日本のヘルスケア分野におけるLLMの現状と課題

2025年9月11日



SB Intuitions株式会社 事業戦略部
JaDHA(日本デジタルヘルスアライアンス) 特別顧問
碇﨑 裕晃

# INDEX
目次

# ヘルスケア分野におけるLLMの性能

SB Intuitions

- 既にさまざまなベンチマークが提案されている。
- 一般的には大学生〜大学院生レベルと同等の性能が出ると言われている。

| | 名称 | 概要 | 引用元 |
|---|---|---|---|
| 1 | MedQA | いわゆる医師国家試験のデータセット | https://github.com/jind11/MedQA |
| 2 | PubMedQA | PubMedから作成された質問応答データセット | https://github.com/pubmedqa/pubmedqa |
| 3 | JMED-LLM | 日本語の医療分野におけるデータセット | https://github.com/sociocom/jmed-llm |
| 4 | AnswerCarefully Dataset | 安全性・適切性に特化したインストラクションデータ | https://llmc.nii.ac.jp/answercarefully-dataset/ |

# ヘルスケア分野におけるLLMの性能

SB Intuitions

- 2022年のChatGPT登場以降、LLMの性能は上がり続けている。
- ただし、性能評価は主に医師国家試験のような静的な知識テストに依存してきた。

| 既存のデータセット、<br>ベンチマーク | 実際の臨床現場における業務 |
| --- | --- |
| モデルの基礎的な能力を評価するためには<br>極めて重要。<br><br>一方で実際の臨床現場における<br>動的なプロセスには対応できていない。 | 実際の臨床現場、実際の業務では、<br>医師や医療従事者は<br>患者との継続的なやり取りや検査結果に基づき、仮説を形成し修正し続ける。 |

# 臨床現場の業務プロセス

SB Intuitions

- 臨床現場における全ての業務は繋がっている。
- それぞれのSTEPで医師や医療従事者は、状況に応じて対応を臨機応変に変化させる。

# ヘルスケア分野における動的プロセスの評価

SB Intuitions

- 静的な性能しか測ることができないという課題感に対しては、GoogleのAMIEやMicrosoftのMAI-DxOに関する先行研究がアプローチしている。
- これらの研究は、静的なタスクから会話的対話や逐次的な情報収集といった動的プロセスへの評価の移行を進めた点で、大きな成果を出している。
- しかし依然として、臨床現場の業務プロセス全体の中で、どの具体的な認知課題がモデルの弱点であるのかを明確には示していない。

| | 名称 | 概要 | 引用元 |
|---|---|---|---|
| 1 | AMIE（Google） | 医療面接と診断を支援するための対話型AIシステム | https://research.google/blog/amie-a-research-ai-system-for-diagnostic-medical-reasoning-and-conversations/ |
| 2 | MAI-DxO（Microsoft） | 診断を支援するためのAIツール | https://microsoft.ai/news/the-path-to-medical-superintelligence/ |

# INDEX
目次

# 論文紹介

SB Intuitions

- A Novel Framework for Evaluating the Clinical Reasoning Process of Large Language Models: A Comparative Study in Nephrology
  - 引用元： https://www.medrxiv.org/content/10.1101/2025.09.04.25334460v1

**A Novel Framework for Evaluating the Clinical Reasoning Process of Large Language Models: A Comparative Study in Nephrology**

Brief title: Evaluating LLM Clinical Reasoning in Nephrology

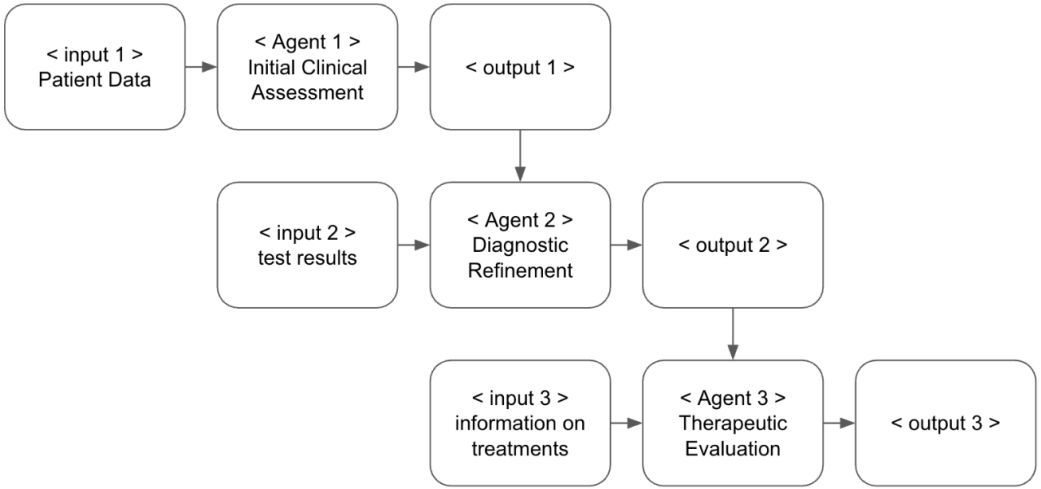
Yuichiro Yano;[1,2] Hiroaki Kakizaki:[3] Hajime Nagasu;[4] Seiji Kishi;[4] Takeo Koshida;[5] Yoshihito Nihei;[5] Akira Hirano;[4] Masaomi Nangaku;[6] Hirotake Mori; [1] Toshio Naito; [1] Mizuki Ohashi;[1] Shoichi Maruyama;[7] Isao Matsui;[8] Yoshitaka Isaka;[8] Yusuke Suzuki;[5]and Naoki Kashihara[4]

# 論文紹介
# 臨床現場の業務プロセスにおけるLLMの性能評価

SB Intuitions

- AIの最終的な出力だけでなく、「臨床推論そのもののプロセス」を専門医が精査した。
- 各症例を「病歴」「検査結果」「その後の臨床経過」の3段階に分けて評価を実施。

**System Overview Diagram of the Multi-Agent Clinical Reasoning Process**



# 論文紹介
# 臨床現場の業務プロセスにおけるLLMの性能評価

SB Intuitions

- 本研究では、4つのモデルを対象に評価を実施した。
- 業務の3段階に合わせてエージェントを構築し、9つの質問を解かせてその性能を評価した。

**Agent 1: Initial Clinical Assessment**

1. 患者の医療上の問題の体系的要約
2. 鑑別診断とその根拠
3. 作業仮説を検証するために必要な追加の質問、または診察の一覧
4. 推奨される診断検査（その目的と期待される結果を含む）

→

**Agent 2: Diagnostic Refinement**

5. 検査結果を解釈し、重要な特徴を強調する
6. この段階でもっとも可能性が高い診断を示す
7. これまでの情報を総合して治療計画を提案する

→

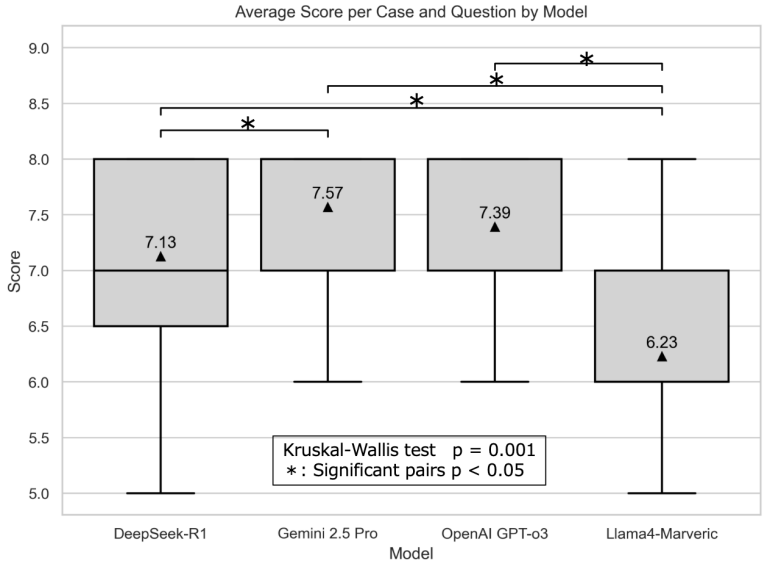
**Agent 3: Therapeutic Evaluation**

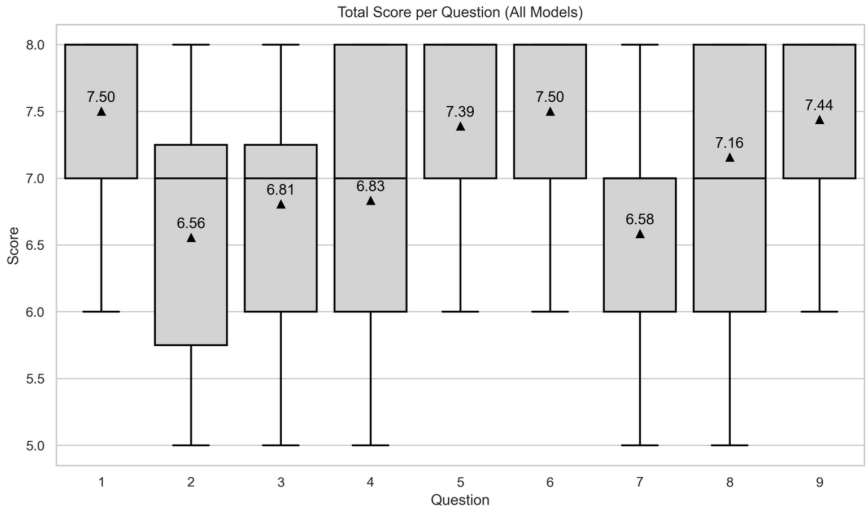
8. 治療の効果や安全性を見守るための方法をまとめる
9. 患者の状態が改善しない場合や予想外に悪化した場合に備えた対応策を提案する

# 論文紹介
# 評価方法と結果の概要

SB Intuitions

- 4名の腎臓専門医が評価し、1名の盲検化された研究者によって集計・解析された。
- 各質問は3点尺度で採点（0 = 誤り、1 = 妥当ではあるが最適ではない、2 = 正しい）



# 論文紹介
# モデル別、質問別の結果

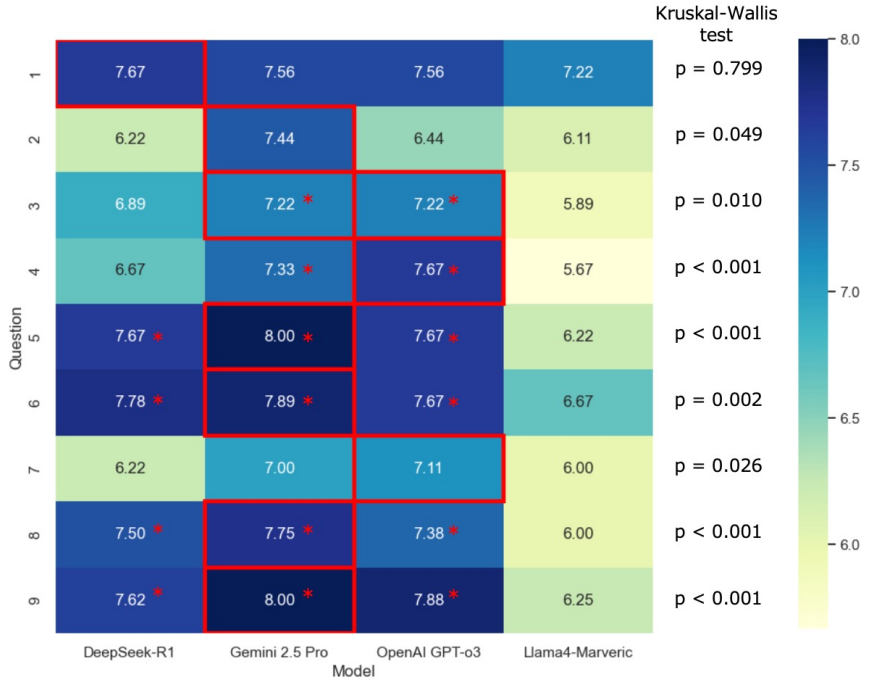
SB Intuitions

- 4つのLLM間で全体的なパフォーマンスに統計的に有意な差が認められた。
- さらにプロセス全体の中でも特に生成AIが弱い部分が明らかになった。

**<STEP 1>**
1. **患者の医療上の問題の体系的要約**
2. **鑑別診断とその根拠**
3. **作業仮説を検証するために必要な追加の質問または診察の一覧**
4. **推奨される診断検査（その目的と期待される結果を含む）**

**<STEP 2>**
5. **検査結果を解釈し、重要な特徴を強調する**
6. **この段階でもっとも可能性が高い診断を示す**
7. **これまでの情報を総合して治療計画を提案する**

**<STEP 3>**
8. **治療の効果や安全性を見守るための方法をまとめる**
9. **患者の状態が改善しない場合や予想外に悪化した場合に備えた対応策を提案する**



# 論文紹介
# 結論、そのほか得られた知見

SB Intuitions

- 一連のプロセスの中で、特にLLMが弱いのは
  「質問2：鑑別診断とその根拠」および「質問7：治療計画」だった。

- 最新のLLMは、情報の要約や、検査結果の解釈といった個別タスクには優れていても、
  最適な介入を計画するような、
  より高度で統合的なスキルには依然として能力のギャップがあることを示している。

- 本研究ではRAGや、ドメイン特化モデルは利用していない。
  ヘルスケアなどの専門性が高い分野においては、
  モデル単体の利用では限界があることが改めて確認できた。

# INDEX
目次

SoftBank

出資 (100%)

# SB Intuitions

- 日本語性能の高い大規模基盤モデルの研究開発
- 日本の文化・商習慣にあった安心安全なAIサービスを提供

**生成AIの開発経験のあるエンジニアがグループ各社から集結**

# 会社概要

SB Intuitions

| | |
|---|---|
| **社名** | SB Intuitions株式会社（英文： SB Intuitions Corp.） |
| **所在地** | 東京都港区海岸一丁目7番1号 |
| **設立年月日** | 2023年3月27日（社名変更日：2023年8月1日） |
| **事業内容** | 日本語に特化した国産の大規模言語モデルの研究開発、生成AIサービスの開発、販売、提供 |
| **代表者** | 代表取締役社長 兼 CEO　丹波 廣寅 |
| **株主** | ソフトバンク株式会社 100% |

# ソフトバンク × AI

SB Intuitions

## AI戦略を積極投資

LLMの内製開発への取り組み、ビジネス応用まで広くカバー

| ソフトバンクグループ | ソフトバンク | SB Intuitions | Gen-AX |
|---|---|---|---|
| **AGI革命を加速する群戦略**<br>AIへの積極投資を実行 | **AGI革命を加速する群戦略**<br>最先端のAI計算基盤を世界初導入 | **国産LLMの自社開発**<br>国内最大級モデルの実現 | 生成AI時代における**トランスフォーメーションを支援** |
| クリスタル・インテリジェンス | 国内最大級のAI計算機版の整備 | 日本語に特化した大規模言語モデルの構築 | 生成AIの社会実装（SaaS/コンサル） |
|  |  | Sarashina | X-Boost |

連携　連携　連携

# ソフトバンク社の国内最大級のAI計算基盤構築

SB Intuitions

- 1兆パラメータの基盤モデル開発に向け、国内最大級※のAI計算基盤を構築

FY23 10月 稼働開始

## NVIDIA DGX™ A100

約2,000GPU (0.7EFLOPs)


FY24 10月 稼働開始

## NVIDIA DGX™ H100

約4,000GPU (4EFLOPs)


FY25 7月 稼働開始

世界初導入

## NVIDIA DGX™ B200

約4,000GPU (9EFLOPs)


累計 約2,000GPU → 約6,000GPU → 約10,000GPU

※現在稼働している「NVIDIA Blackwell GPU」を搭載した「NVIDIA DGX SuperPOD」（AI計算基盤）として世界最大。ソフトバンク調べ（2025年7月23日時点）

# 日本語国産LLM(Sarashina)

Sarashina

**SB Intuitions株式会社が開発するAIモデルシリーズ。**

**日本語に特化したモデル**として開発されており、
日本語の理解や生成において高い性能を発揮する。

## 日本語で生成AIを作る意義

### 言語特性への対応


『漢字・ひらがな・カタカナ共存』や
『敬語・方言のニュアンス』を学習し、
汎用モデルでは捉えられない
日本語の複雑さを自然に反映できる。

### 文化・文脈への理解

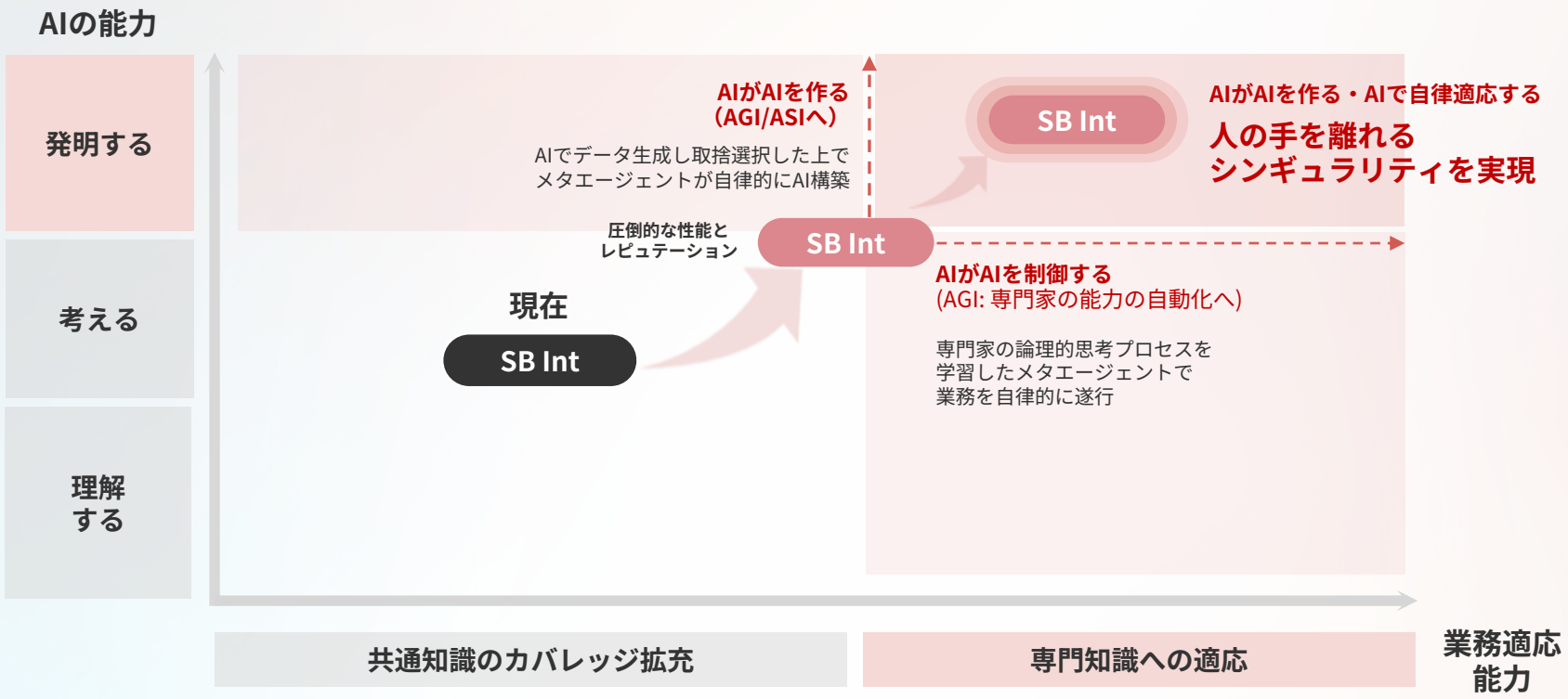
日本語の表現には、
文化的背景や暗黙の了解が
（「忖度」「空気を読む」など）密接に関わる。
こうしたニュアンスをより正確に反映できる。

### 規制・倫理への対応

各国の法規制や倫理観は異なる。
日本語モデルを国内で開発することで、
日本の法律や社会的規範に沿った
運用が可能になる。

# 実現するポジション

SB Intuitions



# あらゆる産業の競争力の源泉となり世の中を成長させる


大規模モデル

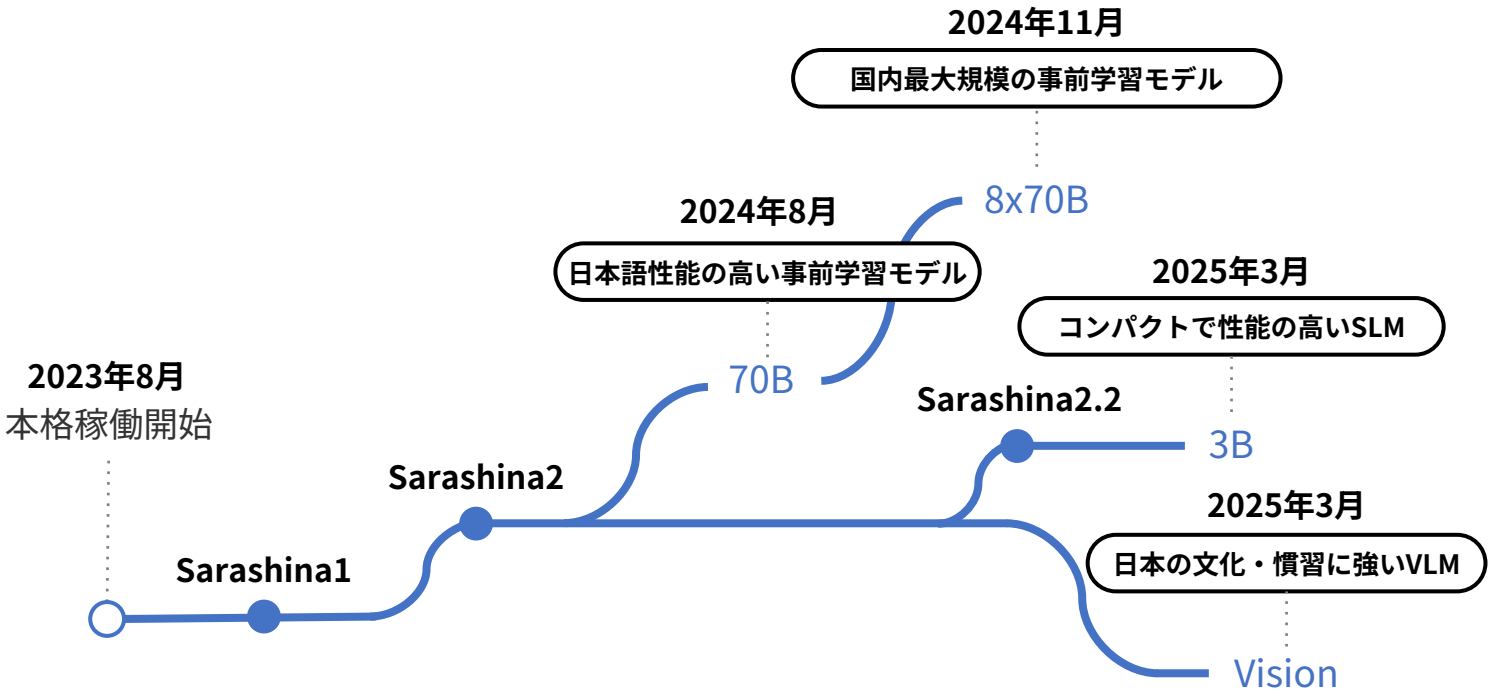
AIプラットフォーム

国内最大規模のAI計算基盤

# SB Intuitionsの国産LLM開発の歩み

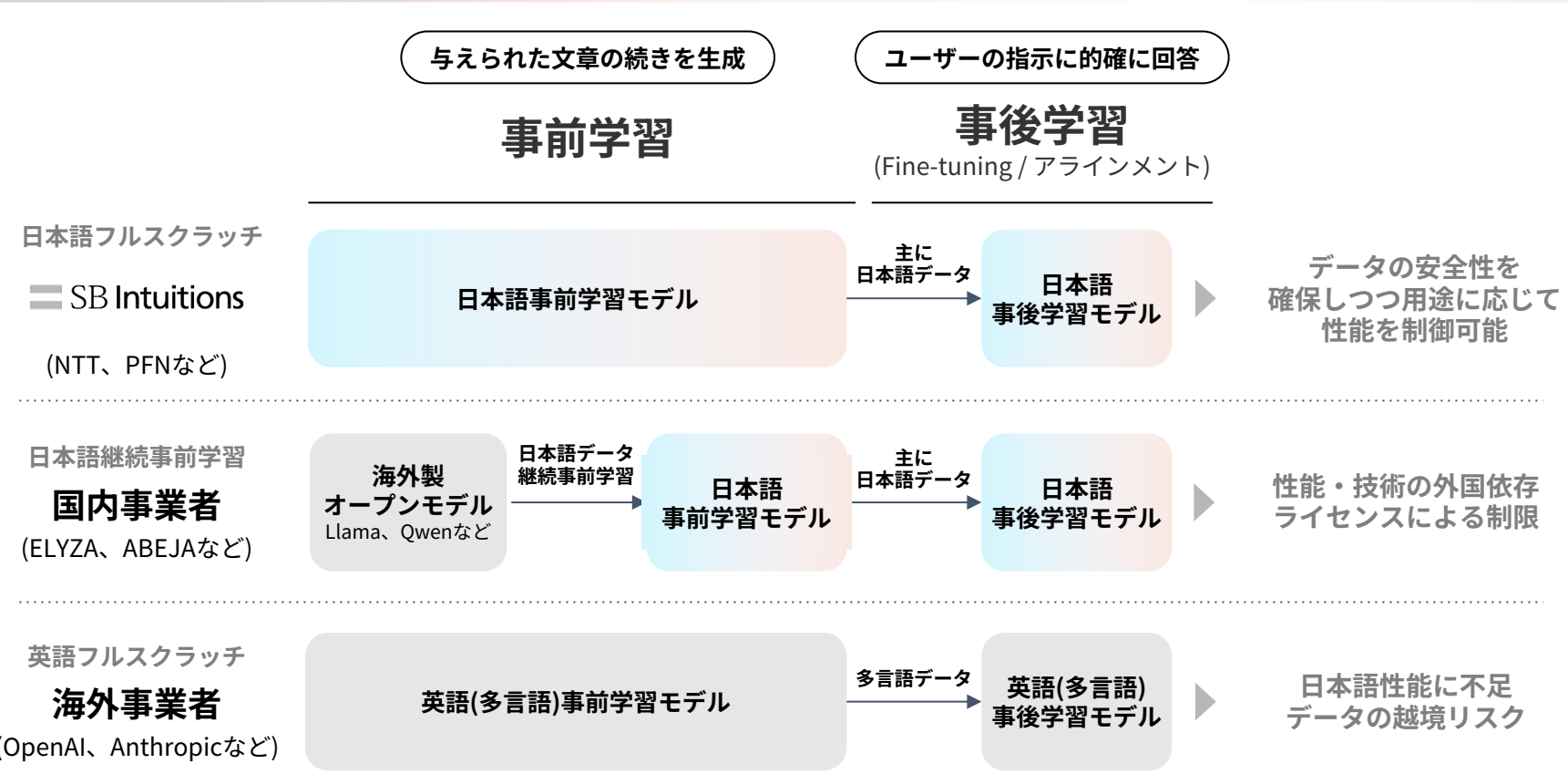
SB Intuitions

- Sarashina1、Sarashina2、Sarashina2.2シリーズを開発し、オープンモデルで公開



# 豊富な計算資源を活かした基盤モデル構築アプローチ

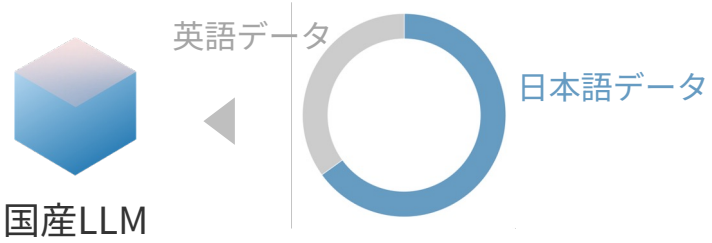
SB Intuitions

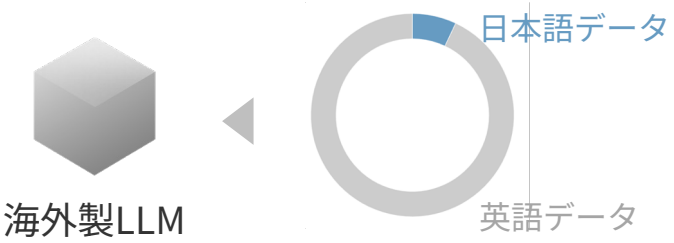


# 学習するデータによる性能の違い

SB Intuitions

- 海外製モデルは学習データのほとんどが英語データであり、
  日本語データの割合は数パーセントに留まる
- 学習させるデータ（日本語/英語）によって、モデルの性能に差異が生じる※




※割合は概念図であり、正確な割合を反映するものではありません。

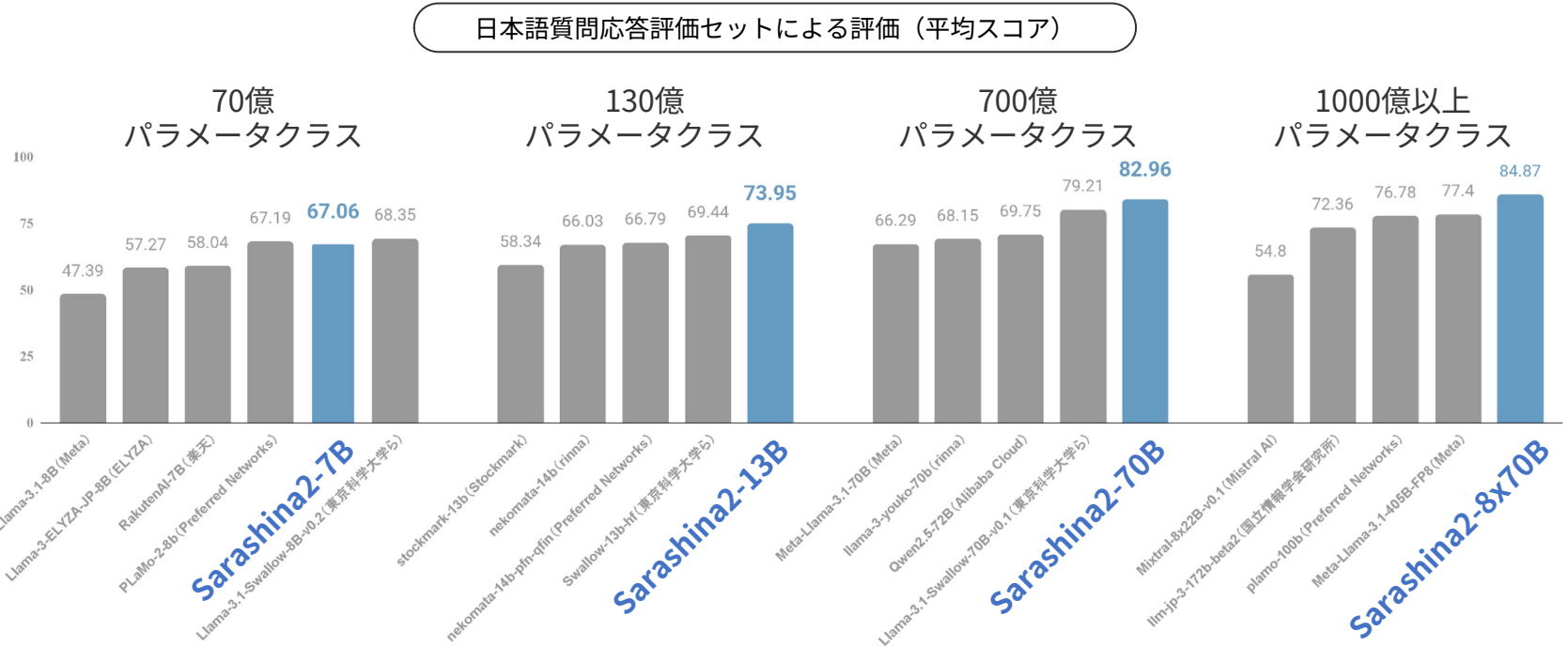
| | |
|---|---|
| 日本語データの学習効果が顕著 | 日本の知識に関する質問応答、日英翻訳 |
| 英語データからでも学習可能 | 日本語の一般教養、算術推論、コード生成 |

※第261回自然言語処理研究発表会での東京科学大学と産業技術総合研究所に所属する研究者らが発表した「LLMに日本語テキストを学習させる意義」より

# 事前学習モデル｜日本語性能に強みを持つ

SB Intuitions

- 「Sarashina2」が各パラメータクラスにおいてトップクラスの日本語性能を記録

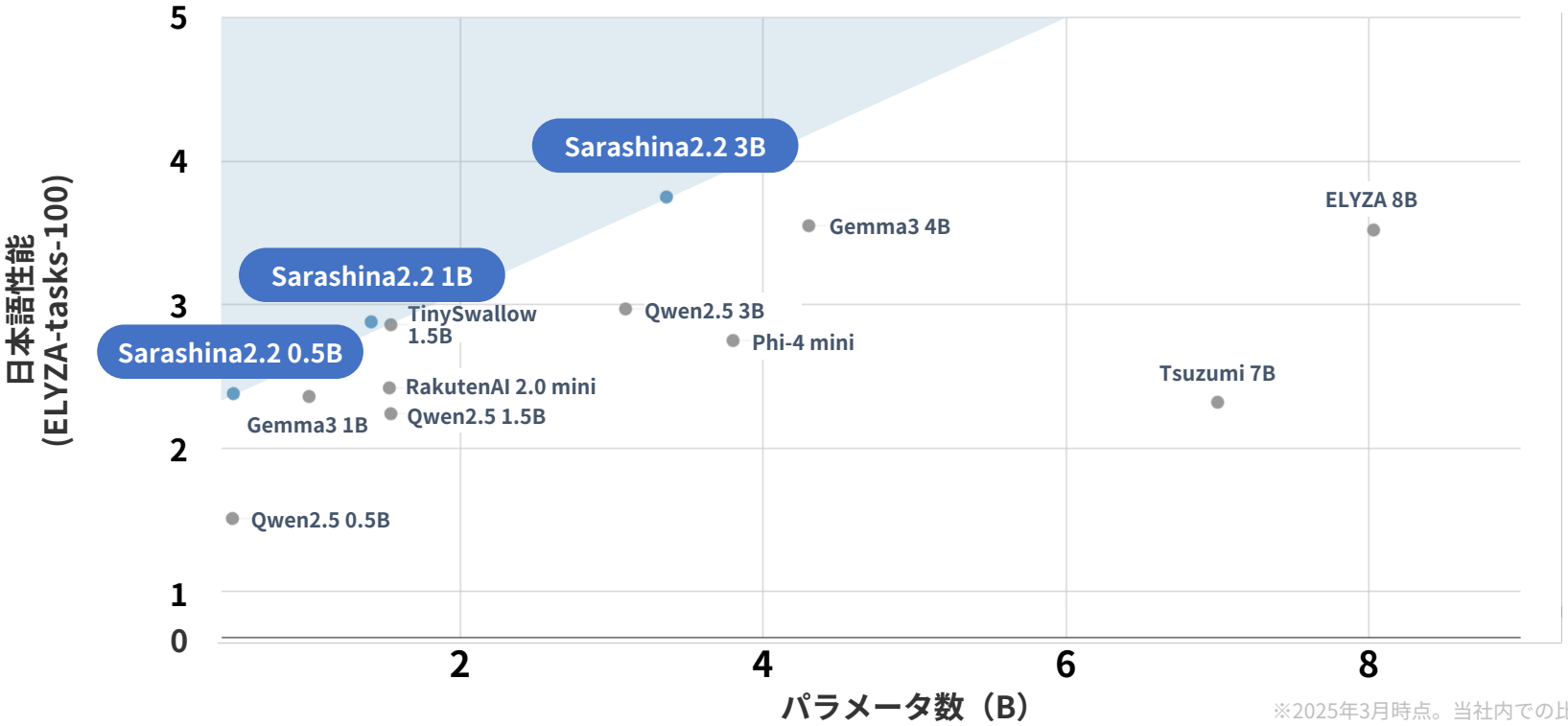



※2025年3月時点。当社内での比較による。

# 事後学習モデル｜小規模で高い性能を持つSLM

SB Intuitions

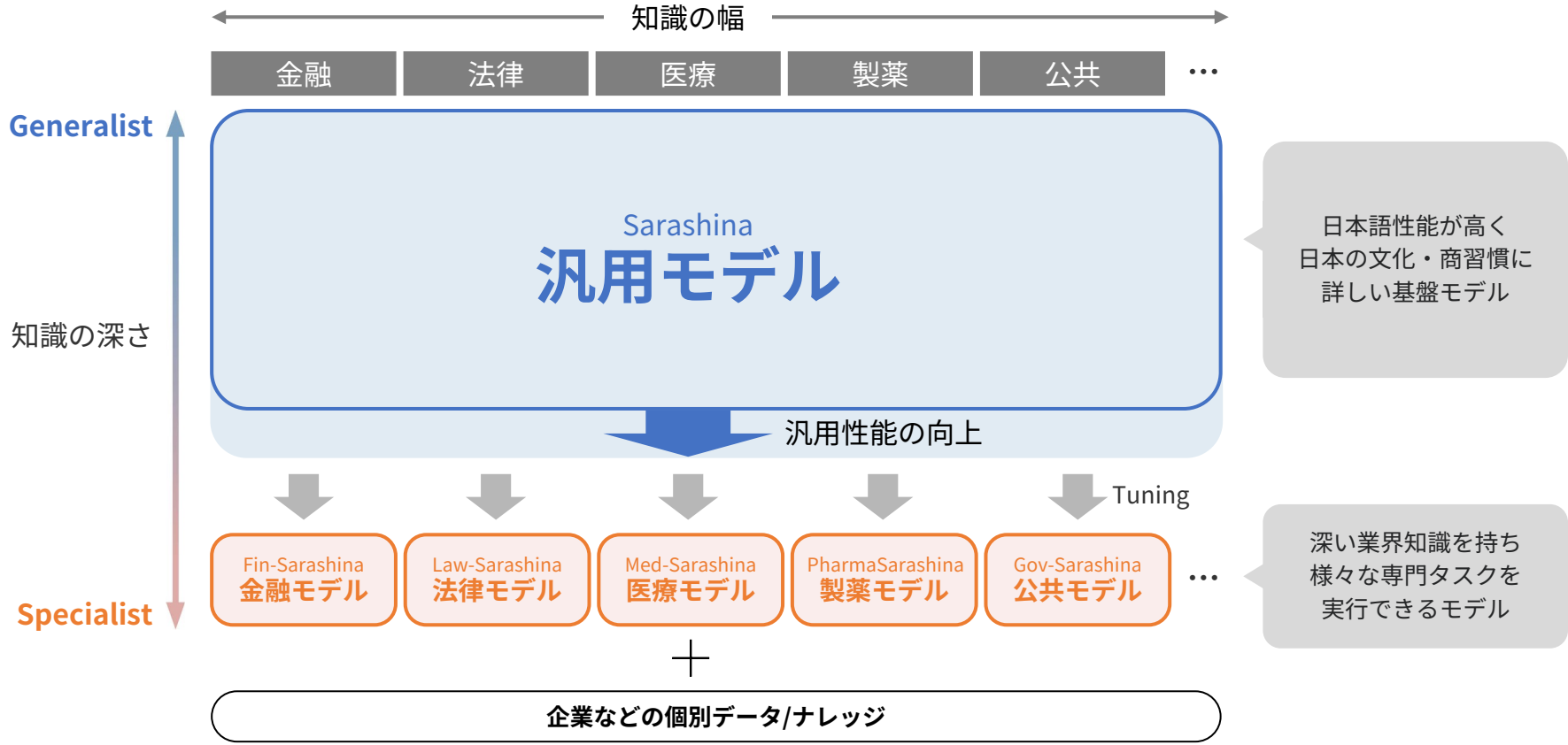
- 「Sarashina2.2」が同じパラメータ規模で最も高い性能をマーク




※2025年3月時点。当社内での比較による。

# 専門領域もカバーできる性能に向けて

SB Intuitions



# モデルをドメインに適応させる方法

SB Intuitions

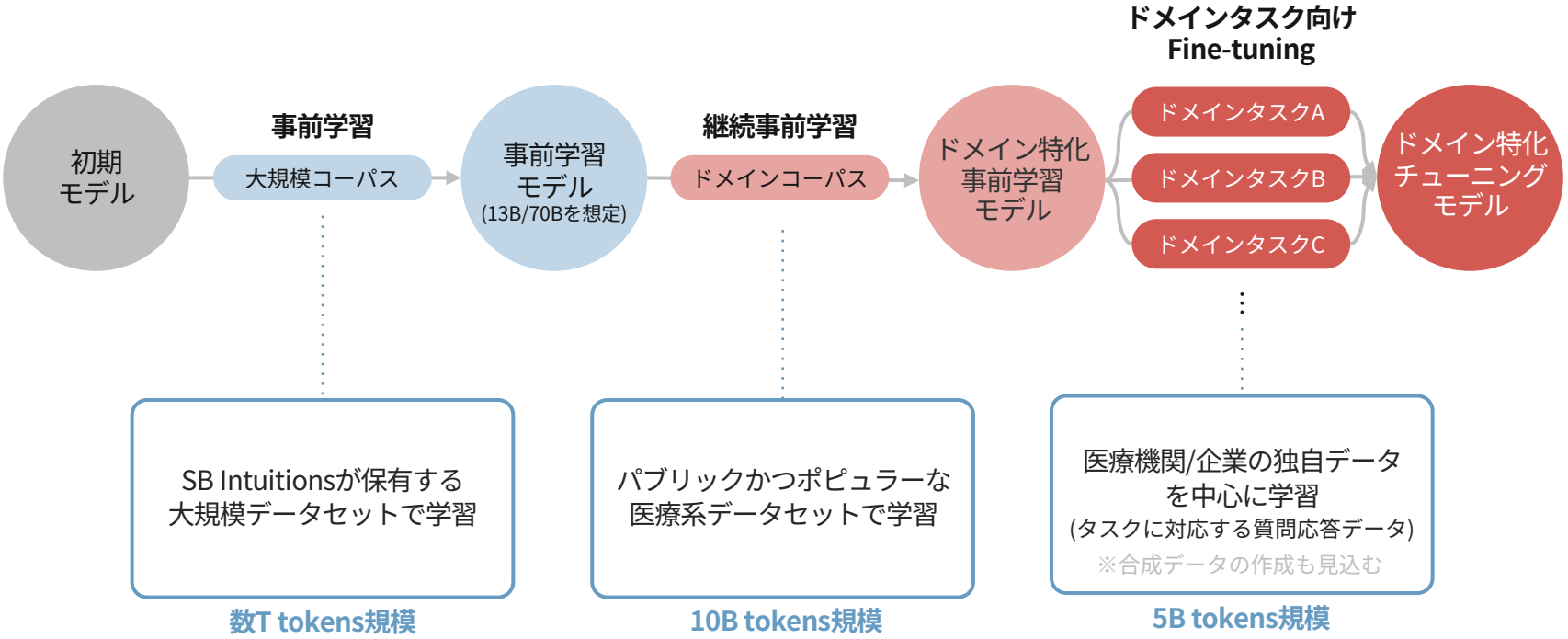
- ドメイン適応には、以下のアプローチをうまく組み合わせることが必要となる
- モデル単体の議論ではなく、RAGやPromptingも視野に入れた総合力で性能を測る

| | |
|---|---|
| Prompt Engineering | • モデルの出力をより正確に制御するために、ドメイン固有の適切な指示やコンテキストを与える手法。<br>• 新しいデータの追加やモデルの再学習を必要とせず、ドメイン固有のタスクに対する性能を向上させられる最も簡単な方法。 |
| 外部知識の活用（RAG） | • モデルが保持していない外部知識を活用することで、モデルの回答の精度や適用範囲を広げる方法。<br>• モデルのトレーニングが最新情報や特定のドメインに限られていない場合、外部ソースからリアルタイムで知識を取得してより正確な回答をすることが可能。 |
| Instruction Fine-tuning | • モデルが特定の命令に従ってより正確に動作するようにFine-tuningする方法。<br>• 特定のタスクやドメインに関するデータセットを使用して、モデルを学習することで、特定のタスクやドメインに適応した回答を行うことが可能となる。 |
| 継続事前学習 | • 既存のbaseモデルに対して、特定のドメインデータセットを用いて、さらに事前学習を行う方法。<br>• 大量の一般的な知識を持つモデルに対して、特定のドメインのテキストデータを用いて再学習を行うことで、そのドメインにおける専門知識や理解を向上させる。 |

コスト

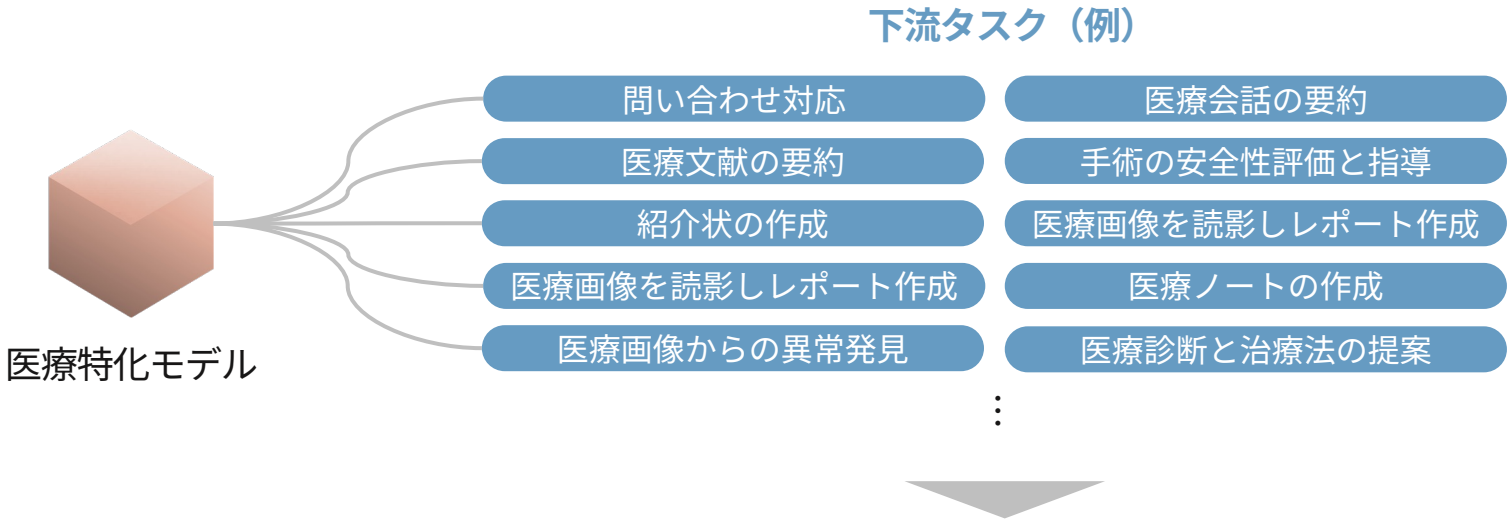
# ドメイン特化モデルの構築の流れ

SB Intuitions



# 有効性が高い下流タスクの選定

SB Intuitions




効果的な下流タスクの見極めと有効性評価が必要

# 今後の展開スケジュール

SB Intuitions

- Sarashina miniから先行して商用開始へ

**2025年3月**

商用モデル完成
Sarashina mini
(700億パラメータ)


**2025年6月**

最終確認
(社内トライアル)

**2025年 秋**

商用提供開始

**2025年度以降**

大規模・高性能
モデルの構築

高性能なSarashina miniを
組み合わせてさらに大規模・高性能化

# 直感を、知性へ