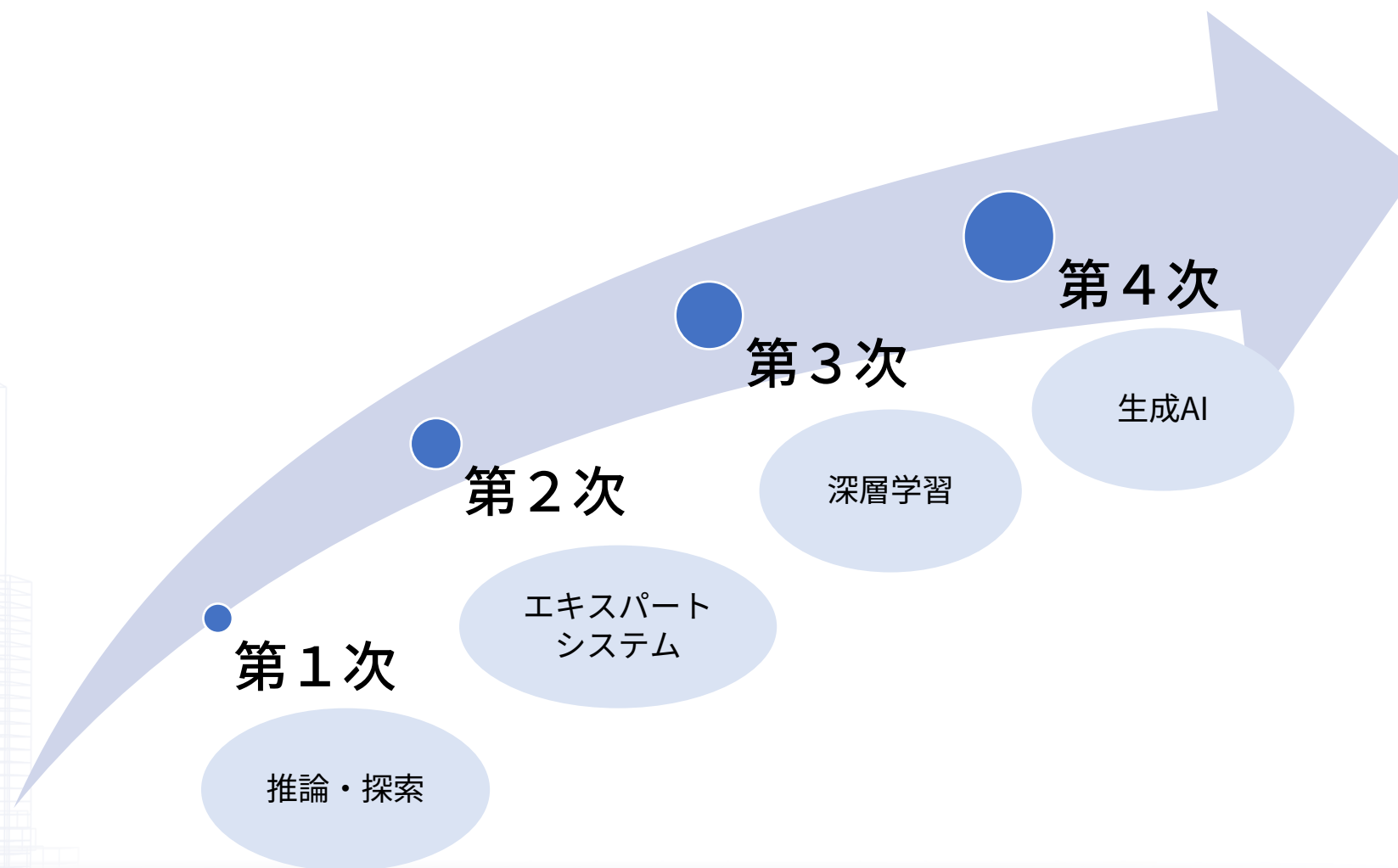


# 医療分野におけるLLMの現状と 応用可能性について

～アカデミア／病院の立場からのLLMへの期待と課題～

# 第4次AIブームの到来



# 第4次AIブームの到来

## ChatGPTが医師国家試験「合格」も、診療利用に不向きな理由

🔒 有料記事 ⓘ ChatGPT

枝松佑樹 2023年6月15日 17時30分

<https://www.asahi.com/articles/ASR6H3Q13R6GUTFL00H.html>

## 芥川賞作「ChatGPTなど駆使」「5%は生成AIの文章そのまま」 九段理江さん「東京都同情塔」

🕒 2024年01月18日 09時24分 公開

[岡田有花, ITmedia]

<https://www.itmedia.co.jp/news/articles/2401/18/news090.html>

## 生成AIが司法試験「合格水準」 東大発新興、一部科目で「GPT-4」ベースに独自開発

法務・ガバナンス [+ フォローする](#)

2023年6月11日 11:00 [会員限定記事]

<https://www.nikkei.com/article/DGXZQOUC317WP0R30C23A5000000/>

生成AIとは何か、どこまで進んでいるのか、医療領域でも使えるのか、どんな課題はあるのか

# 本日の内容

## 話すこと

- LLMの概要
- 医療分野におけるLLMの研究紹介
- 医療応用に向けたLLMへの期待と課題

## 話さないこと

- 自然言語処理全般の説明
- LLMのアーキテクチャや数式的な説明

# 目次

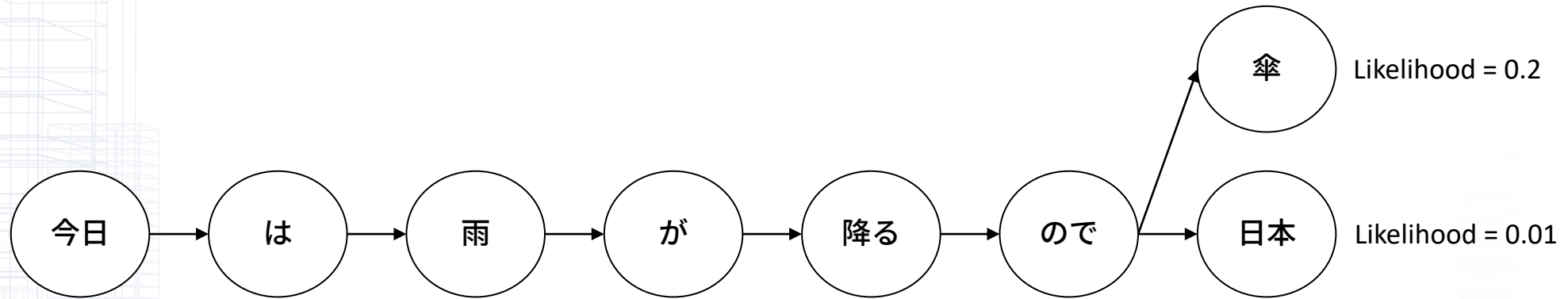
- LLMの概要
- 医療分野におけるLLMの研究紹介
- 医療応用に向けたLLMへの期待と課題

# 目次

- LLMの概要
- 医療分野におけるLLMの研究紹介
- 医療応用に向けたLLMへの期待と課題

# 言語モデル (Language Model; LM)

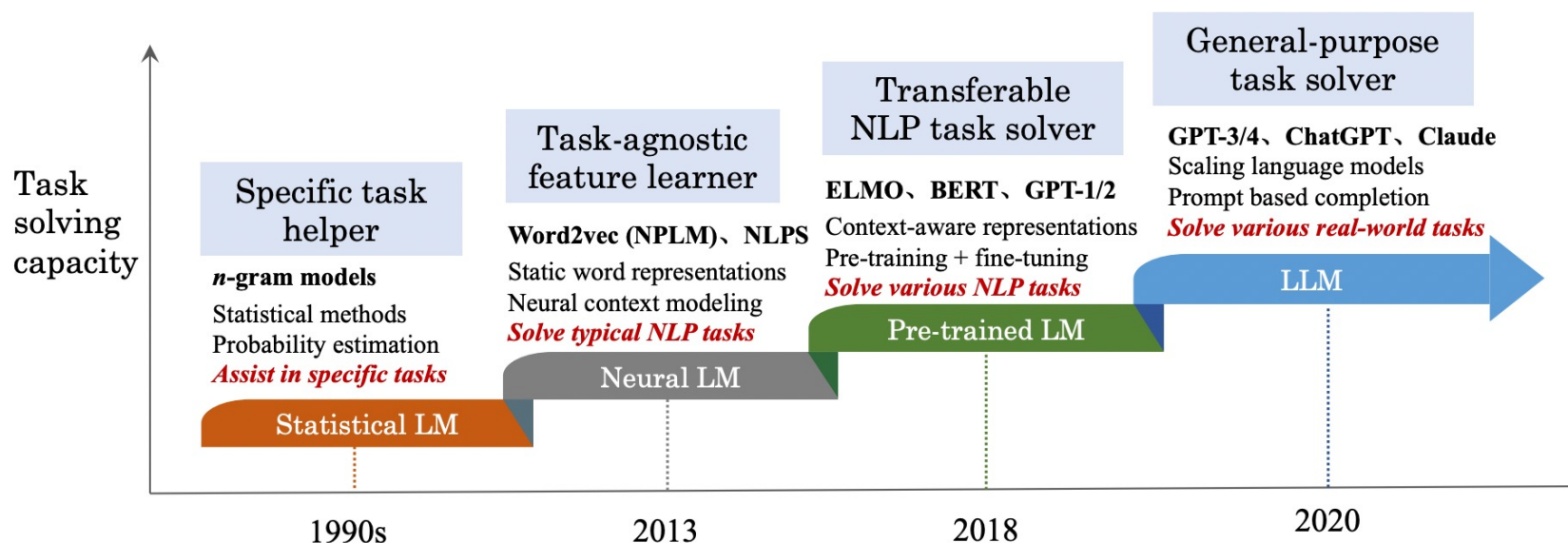
- 言語モデル (Language Model) とは、ある単語系列を入力として、その系列の尤もらしさ (Likelihood) を出力する確率分布モデルとして定義される
- 近年は「言語理解・言語知識を持った機械学習モデル」という意味で使用されることも多く、LLMの文脈でもこの意味で理解するほうが自然



言語モデルの直感的な説明

# 言語モデルの発展

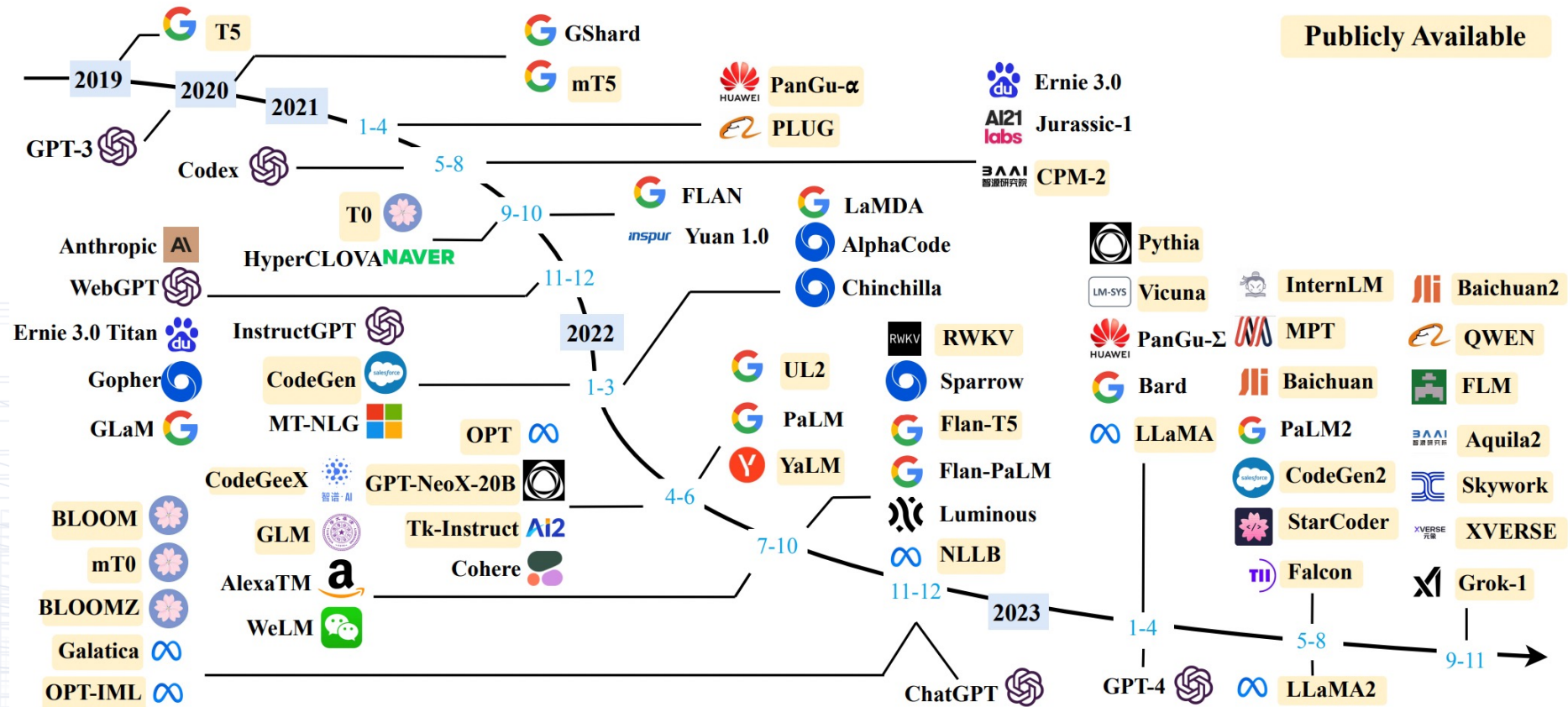
- 自然言語処理の分野でも大きなパラダイムシフトが起きている
- 教師あり学習・ファインチューニングからプロンプト調整の世界に
- 解けるタスクの自由度についても、格段に広がってきている



Zhao et al. 2023



# 大規模言語モデルの発展



Zhao et al. 2023

# Self-supervised learning (自己教師あり学習)

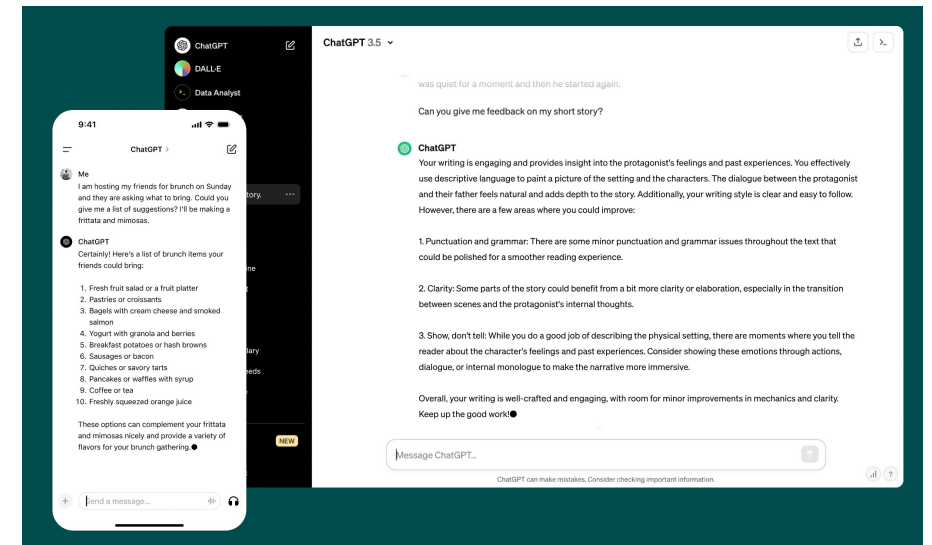
*“Self-supervised learning could lead to the creation of AI that’s more humanlike in its reasoning” (Yoshua Bengio and Yann LeCun)*

<https://venturebeat.com/ai/yann-lecun-and-yoshua-bengio-self-supervised-learning-is-the-key-to-human-level-intelligence/>

- これまでは、専門家がラベル付け作業を行い、学習データを準備するのが一般的だったが、人手によるものなので、データ量に限界があった
- 自己教師あり学習は、人手でラベル付け作業を行うのではなく、データ自身の情報だけでモデルの学習（データの一部をマスクして予測を行うなど）を行い、データの分布や特徴などを獲得する
- 自己教師あり学習の成功により、Web上の大量データだけを使って、汎用的な言語モデルを構築することが可能となった

# ChatGPT

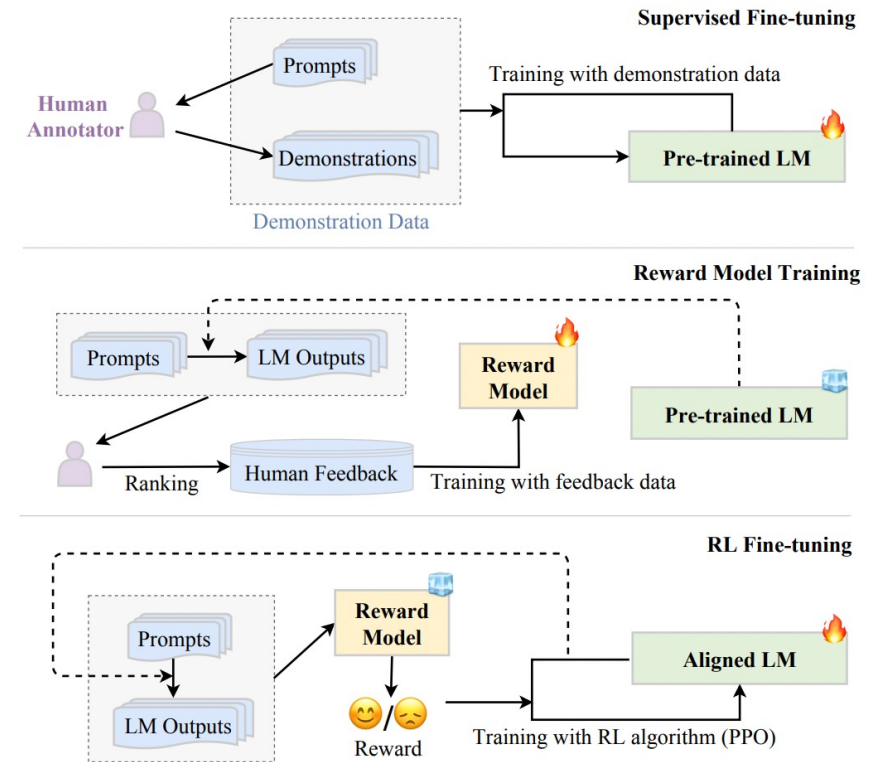
- 言語モデルは単語系列の尤もらしさを最大化するモデルに過ぎず、自然な会話などを行うのは難しいとされており、また、出力結果に有害な情報、役に立たない情報を含むことがしばしばあった
- Instruction tuning (Ouyang et al. 2022) と呼ばれる手法を取り入れ、ユーザーの欲しい情報に応じて適切な出力を調整（アラインメント）することで、ユーザーとの自然な会話・リクエストに応じた出力が可能となった



<https://openai.com/chatgpt>

# Reinforcement Learning from Human Feedback (RLHF)

- ユーザーのリクエストにアラインメントさせる方法として行ったのがRLHFと呼ばれる強化学習の手法の一つ
- 学習の流れとしては
  1. プロンプトを与え、人間のラベラーの出力を使用してLLMを教師あり学習
  2. プロンプトに対するLLMの出力を複数集め、人間がその出力をランク付けする。このランク付きデータを用いて、報酬モデルを学習させる（どの出力が望ましいのかを学習する）
  3. プロンプトの出力から報酬を生成し、報酬を最大化するよう強化学習を行う



Zhao et al. 2023

## まとめ

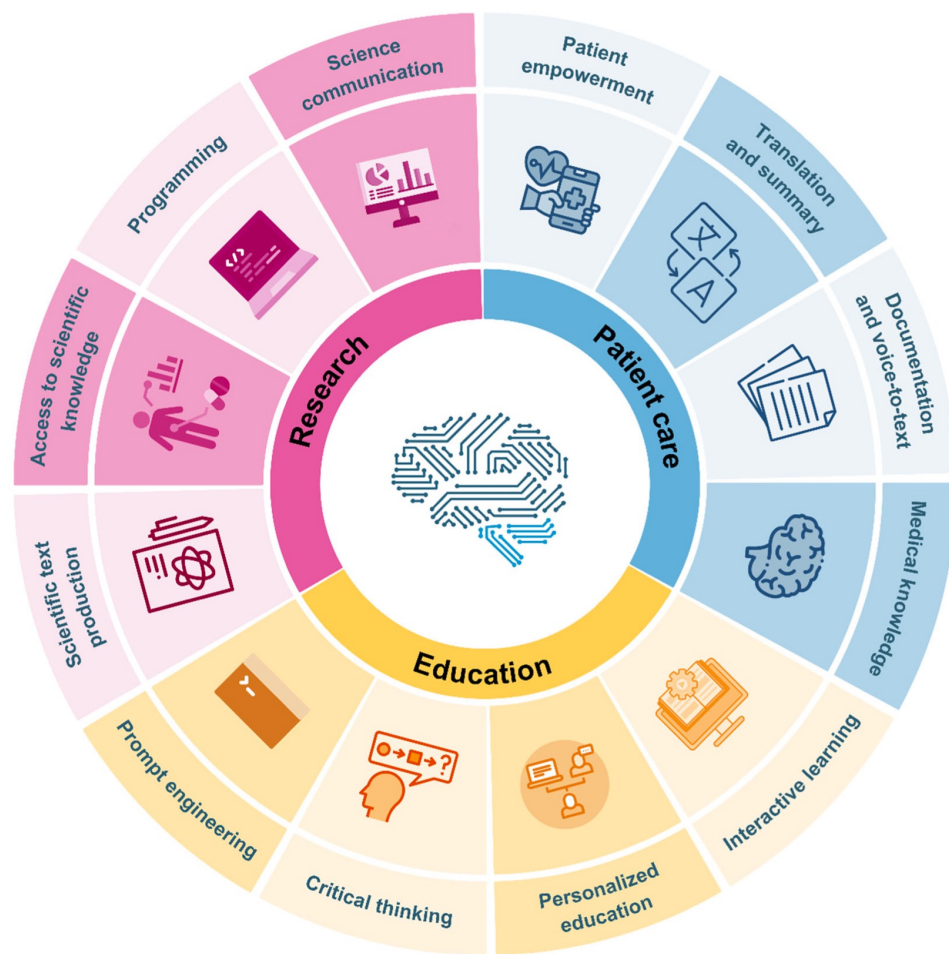
- 教師あり学習はタスクに応じて専門家がラベル付け作業を行い、学習データを準備するのが一般的だったが、人手による作業を伴うものなので、データ量の準備に限界があった
- 自己教師あり学習により、汎用的な言語知識を持つ言語モデルの構築され、タスク毎に少量のデータでファインチューニングする手法が確立された
- 言語モデル、学習データの大規模化により、ファインチューニングを行わず、プロンプトの入力だけで期待する出力を得ることが可能となってきた
- モデルのアライメントにより、言語モデルの限界を超えたユーザーとの自然な会話、有害な出力の制御などの実用面での課題が改善された

# 目次

- LLMの概要
- 医療分野におけるLLMの研究紹介
- 医療応用に向けたLLMへの期待と課題

# 医学領域におけるLLMの応用

1. Education
2. Patient Care
3. Research



Clusmann et al. 2023

# Kung et al. 2023

- ChatGPTに米国の医師資格試験 United States Medical Licensing Exam (USMLE) を解かせた論文
- Step毎の試験を解かせ、いずれの試験でも合格水準の60%に近い水準を達成
  - Step 1（主に基礎科学・薬学・病態生理学などを2年専攻した医学生が受ける試験）
  - Step 2CK（さらに研修を通じて臨床推論・臨床医学を専攻した4年次の医学生が受ける試験）
  - Step 3（さらに卒後教育を終えた後の医師が受ける試験）



# Kung et al. 2023

- ChatGPTは生物学医学の分野に特化したPubMedGPTよりも高い性能であった
- Step 1の精度は最も低く、Step 3が最も高いという結果は実際の受験者の感覚に沿うものであった
- ChatGPTは多くの問題で有用な洞察・論理的な説明（例：副腎皮質機能亢進症により骨破骨細胞の活動が増加した結果、カルシウム吸収が減少し、骨密度が減少して骨折リスクが増加する）を出力できており、将来の医学教育の有効性を示唆している

A

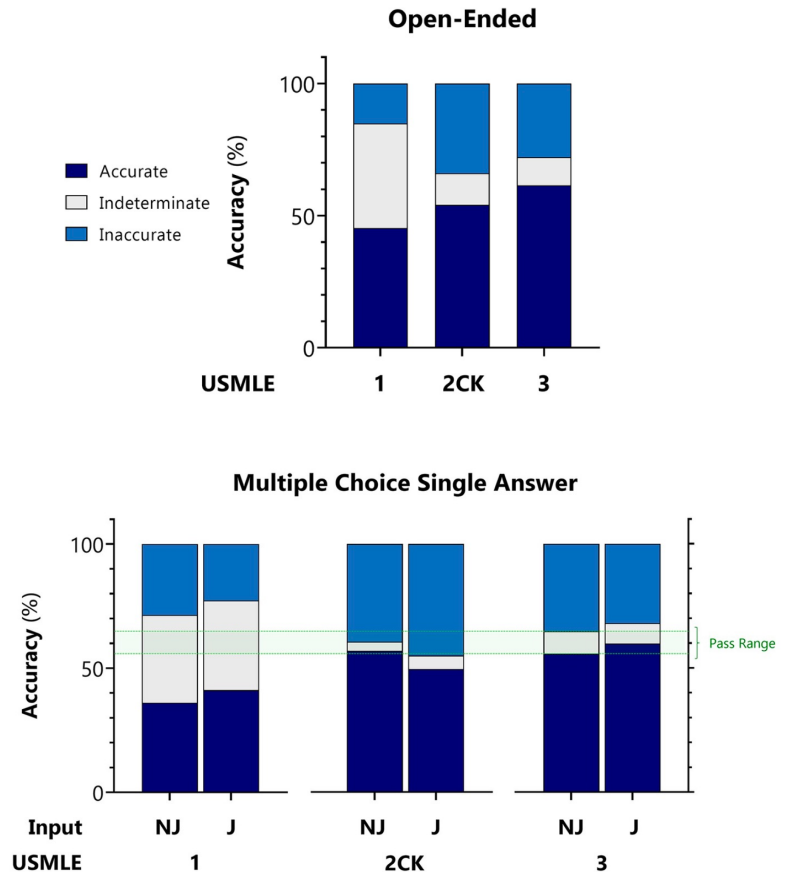
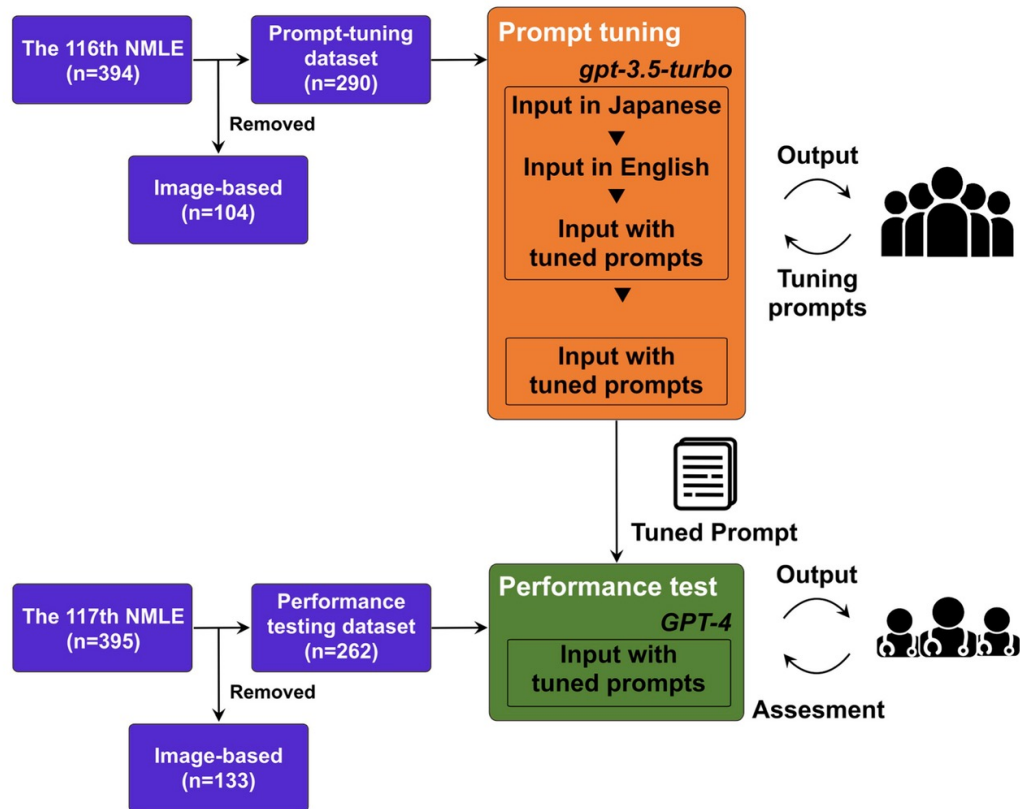


Fig 2. Accuracy of ChatGPT on USMLE. For USMLE Steps 1, 2CK, and 3, AI outputs were adjudicated to be accurate, inaccurate, or indeterminate based on the ACI scoring system provided in S2 Data. A: Accuracy distribution for inputs encoded as open-ended questions. B: Accuracy distribution for inputs encoded as multiple choice single answer without (MC-NJ) or with forced justification (MC-J).

# Tanaka et al. 2024



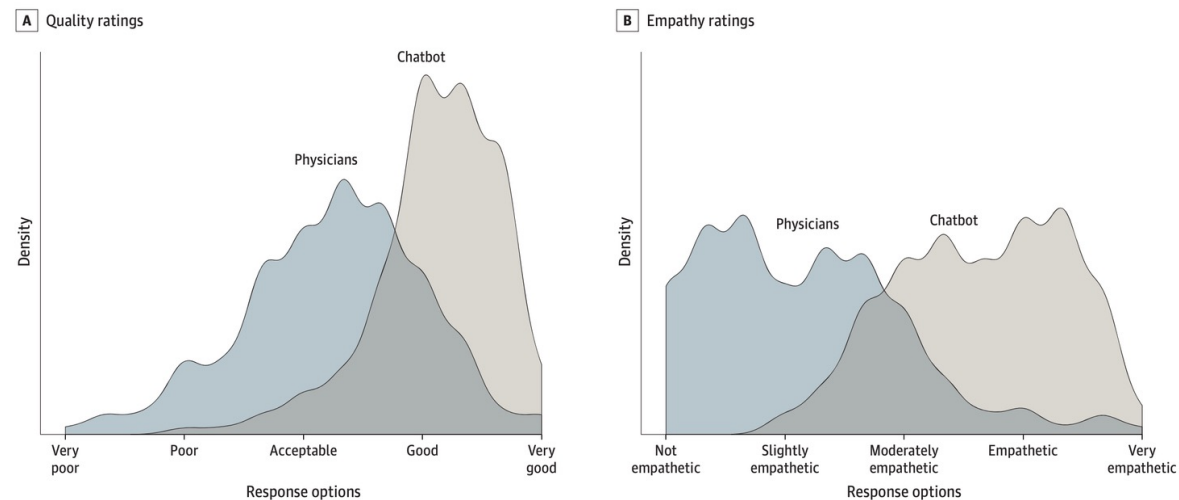
- ChatGPTに日本の医師国家試験を解かせて、合格ラインを達成
- 日本語を英語に翻訳し、簡単な要約などを行うようプロンプトをチューニング
- GPT-4の訓練終了後の試験（117回；2023年2月）を用いて性能を評価
- 医学知識の不足だけではなく、日本の医療制度の問題、数学的な計算問題などのエラーがあった
- 不適切な回答の根拠では、もっともらしい医学知識に基づく説明（ハルシネーション）もあった

## Suchman et al. 2023

- ChatGPTを使用して、米国消化器病学会の多肢選択式の自己評価テストに取り組んだ論文
- 合格基準70%に対し、GPT-3.5で65.1%、GPT-4で62.4%の正答率で、いずれも不合格の判定であり、これまでの肯定的な報告とは異なり、消化器学の医学教育において慎重な議論が必要であることを報告した
- 原因として、ChatGPTが最新のガイドラインによって訓練されていないこと、有料の雑誌へのアクセスが制限されていたことにより、消化器学の領域の最新の知識を十分に獲得できていないとしている

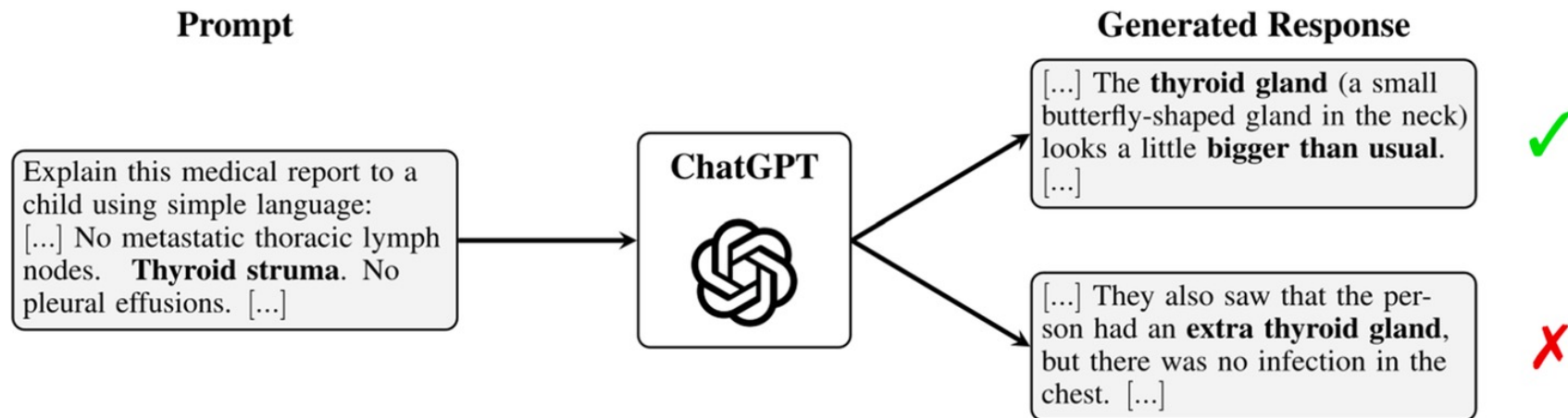
# Ayers et al. 2023

- ChatGPTを用いて一般の患者への回答の品質・内容について評価
- ソーシャルメディアに投稿された質問を用いて、医師による回答と比較
- 評価者は78.6%の指標においてChatGPTの回答のほうが適切と評価した
- 回答が共感的な内容の評価についても、有意にChatGPTの回答のほうが共感的であるという結果であった



# Jeblick et al. 2023

- ChatGPTを用いて、画像診断レポートの平易化の有用性を検証
- 放射線科医が品質・完全性・事実性・一貫性などのいくつかの観点でChatGPTの生成結果を評価したところ、ほとんどの生成結果が完全であり、患者に害を及ぼす可能性はないと判断していた

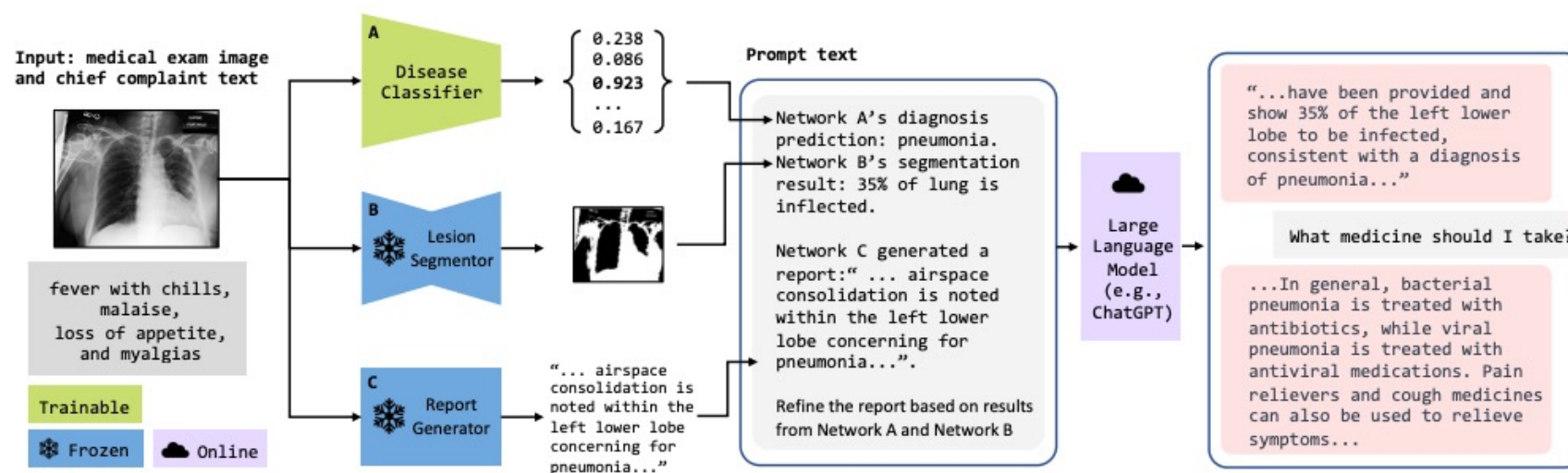


## Sarraju et al. 2023

- ChatGPTが生成する心血管疾患（cardiovascular disease; CVD）の予防のためのアドバイスの有用性・妥当性について評価
- ガイドラインに基づく予防のトピック、臨床経験に基づいて、危険因子のカウンセリング、検査結果、投薬情報などの基本的な予防の概念を扱う25の質問を作成した
- 医師らが作成したCVDに関する25の質問の回答結果を評価した結果、21の回答は適切であるという結果であった
- 回答結果への課題はあるものの、一般的なCVD予防に関する質問に対する患者教育や患者と医師のコミュニケーションを支援するワークフローへの有用性が示唆されたとしている

# Wang et al. 2023

- ChatGPTと画像AIを組み合わせて、胸部単純レントゲン画像の診断支援（CAD）を行った研究
- マルチモーダル（画像・テキスト）を用いた実臨床への有用性を示している



## まとめ

- 専門知識を必要としないタスク（administrative task）や、ユーザープロンプトで情報が提供されるタスクにおいては、優れたパフォーマンスを示すことが多くの研究で示されている（Thirunavukarasu et al. 2023）
- 例えば、退院サマリのような要約・言い換えがメインのタスクでは医療従事者の負担を軽減できる可能性が高いとしている（Patel and Lam 2023）
- その他、ユーザープロンプトで提供できるタスクとして、情報抽出・構造化などが考えられ、例えばフリーテキストの画像診断レポートからの情報抽出などのタスクへの有用性も示されている（Adams et al. 2023）

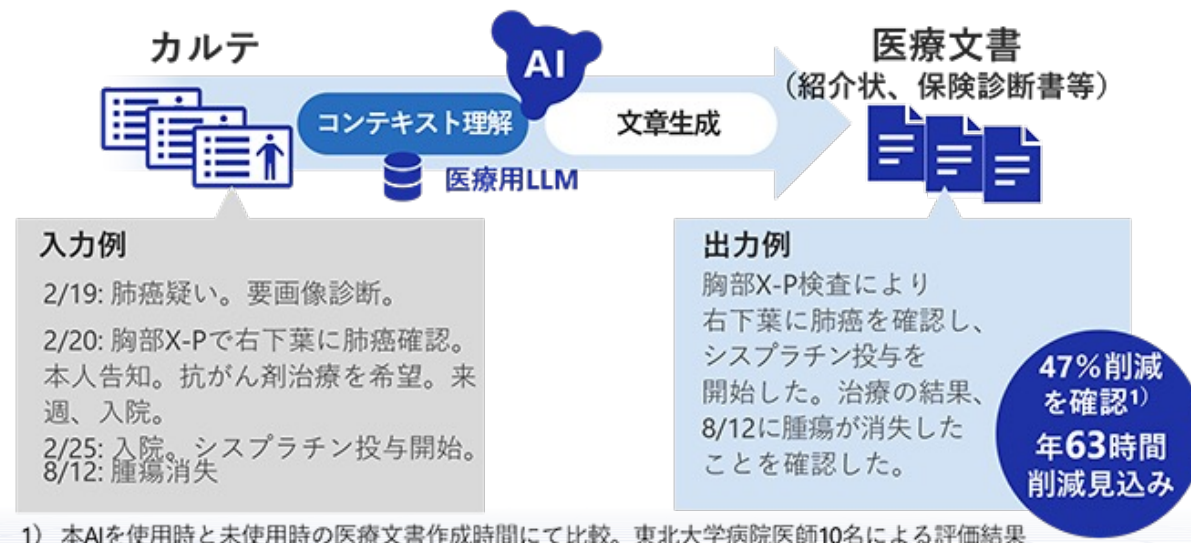


## (参考) 電子カルテからの医療文書作成の検証

- NECと東北大学病院でのLLMを用いた医療文書の自動作成に関する実証実験
- 紹介状などの文章の作成時間が平均47%削減できたとしている

### ■ カルテから医療文書を自動構成

医療用語の意味や治療経過を理解し、多様な医療文書の形式に自動で構成。文書作成時のカルテ読み返し作業が不要に。

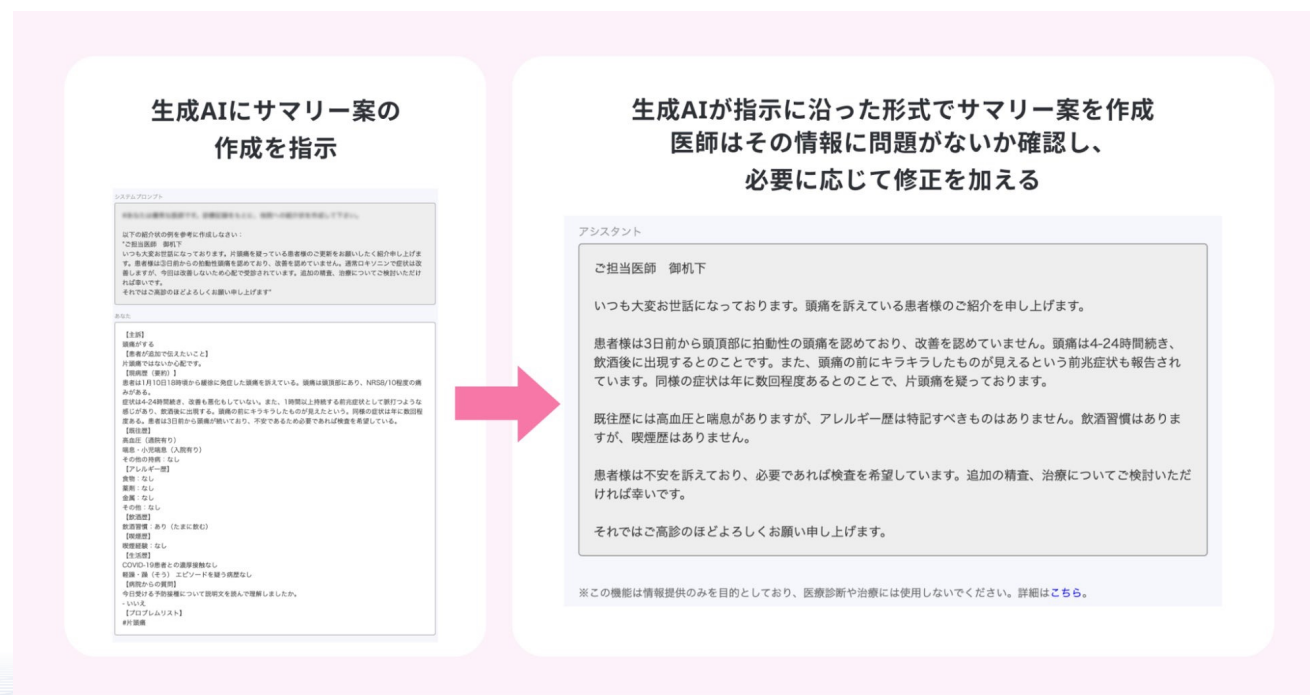


1) 本AIを使用時と未使用時の医療文書作成時間にて比較。東北大学病院医師10名による評価結果

[https://jpn.nec.com/press/202312/20231213\\_01.html](https://jpn.nec.com/press/202312/20231213_01.html)

# (参考) 電子カルテからの医療文書作成の検証

- Ubieと恵寿総合病院でのLLMを用いた医療文書の自動作成に関する実証実験
- 医師の退院時サマリー作成業務を最大1/3にまで短縮できたとしている

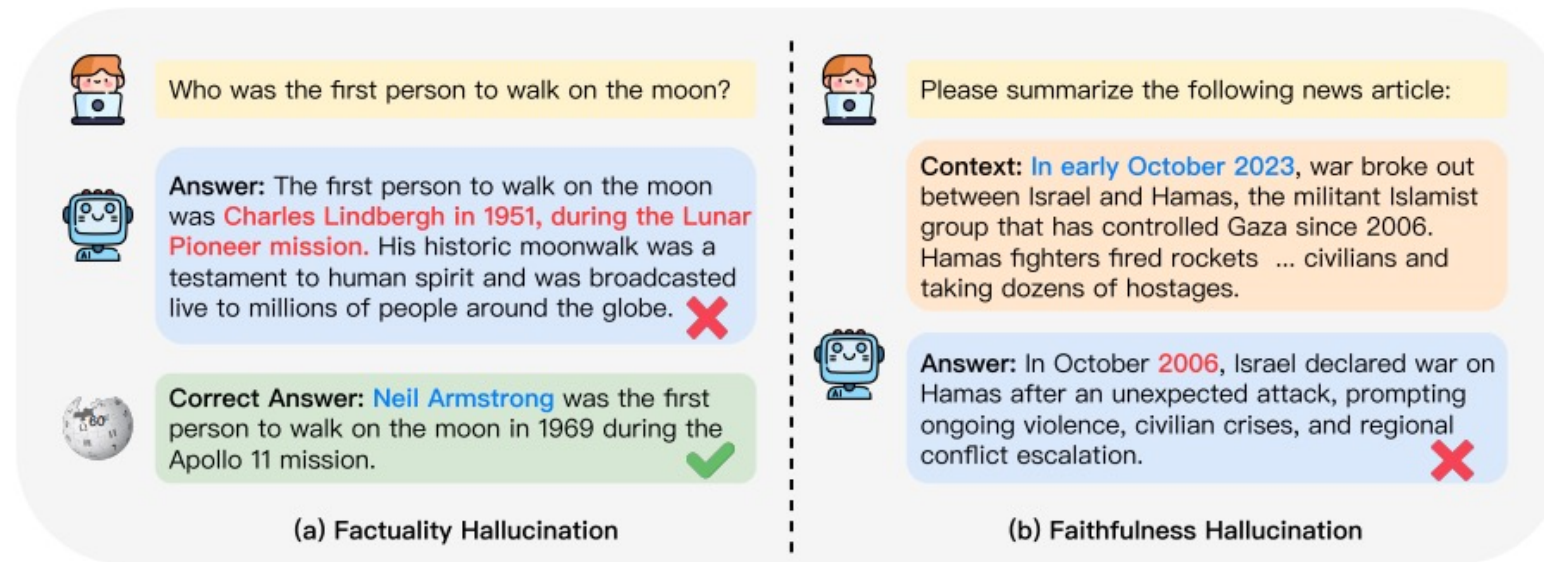


# LLM limitations (Thirunavukarasu et al. 2023)

課題	内容
Recency	LLMで使用する訓練データは特定時点までの情報しか含めることができず、新しい臨床概念・研究の知見などを得ることができない。
Accuracy	自己教師学習は、訓練データの内容を理解しているのではなく、単語間の確率的な関連付けを訓練しているに過ぎない。また、訓練データに使用されるウェブサイトや書籍などの情報の信頼性は保証できない。
Coherence	モデルは与えられた入力をもとに、学習された単語間の関連付けに基づいてテキストを出力しているに過ぎず、存在しない情報もあたかも真実であるかのように出力し得る（ハルシネーション）。
Transparency and interpretability	モデルがどのように回答を生成するのかのプロセスが不明である。また、生成された回答は訓練セットのどの部分に関連しているかは不明である。
Ethical concerns	出力が有害、差別的、攻撃的である可能性があり、また、プライバシーやセキュリティ侵害のリスクなどの懸念もある。

# Hallucination

- Factuality Hallucination
- Faithfulness Hallucination



The diagram is divided into two columns by a vertical dashed line. Each column contains a user question, a model's incorrect answer, and a correct answer.

**(a) Factuality Hallucination**

Question: Who was the first person to walk on the moon?

Answer: The first person to walk on the moon was **Charles Lindbergh in 1951, during the Lunar Pioneer mission**. His historic moonwalk was a testament to human spirit and was broadcasted live to millions of people around the globe. ❌

Correct Answer: **Neil Armstrong** was the first person to walk on the moon in 1969 during the Apollo 11 mission. ✅

**(b) Faithfulness Hallucination**

Question: Please summarize the following news article:

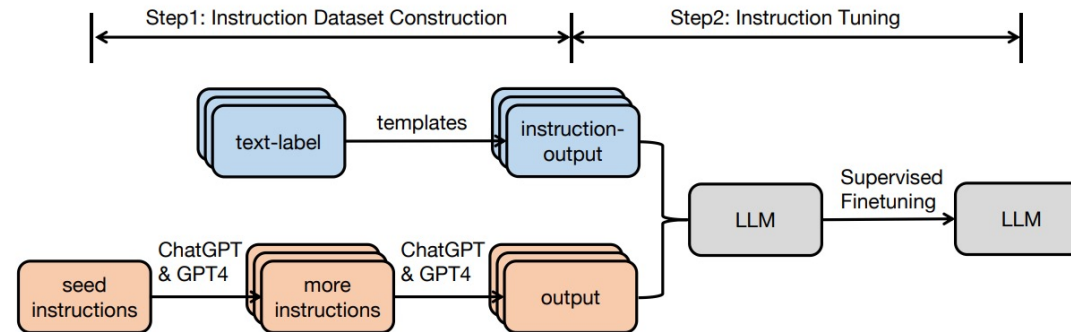
Context: **In early October 2023**, war broke out between Israel and Hamas, the militant Islamist group that has controlled Gaza since 2006. Hamas fighters fired rockets ... civilians and taking dozens of hostages.

Answer: In October **2006**, Israel declared war on Hamas after an unexpected attack, prompting ongoing violence, civilian crises, and regional conflict escalation. ❌

Huang et al. 2023

# Fine-tuning

- 臨床テキスト（診療記録・検査レポートなど）や医療情報文書（ガイドラインや論文など）を使用したドメイン適応
- ただし、多大なリソースとコストを要すること、臨床テキストを含める場合、個人情報の取り扱いの問題から課題も多い



Zhang et al. 2023

# LLMの利用・拡張

- Prompting
- Retrieval Augmented Generation (RAG)
- LLM Agents

# Prompting

- モデルの出力を導くためにユーザーから提供されるテキスト入力（プロンプト）をデザインすること
- 期待する出力を得るためには、適切なプロンプトを入力する必要があり、そのプロンプトを設計するプロンプト・エンジニアリングという分野が急速に発展している

# Wei et al. 2022

- 暗黙の推論プロセスを明示化（Chain of Thought：CoTと呼ばれる一連の間推論ステップを生成）することで、LLMが複雑な推論を行う能力を引き上げることができることを示した研究

## Standard Prompting

### Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

### Model Output

A: The answer is 27. ❌

## Chain-of-Thought Prompting

### Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls.  $5 + 6 = 11$ . The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

### Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had  $23 - 20 = 3$ . They bought 6 more apples, so they have  $3 + 6 = 9$ . The answer is 9. ✅



# Kojima et al. 2022

- マニユアルで中間プロセスを入力 (Few-shot) するのに対し、単に「step by step」を入力に指示するだけのZero-shotのプロンプトでもLLMの性能が向上させることができることを示した研究

(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?  
A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?  
A:

(Output) The answer is 8. ✗

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls.  $5 + 6 = 11$ . The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are  $16 / 2 = 8$  golf balls. Half of the golf balls are blue. So there are  $8 / 2 = 4$  blue golf balls. The answer is 4. ✓

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

(Output) 8 ✗

(d) Zero-shot-CoT (Ours)

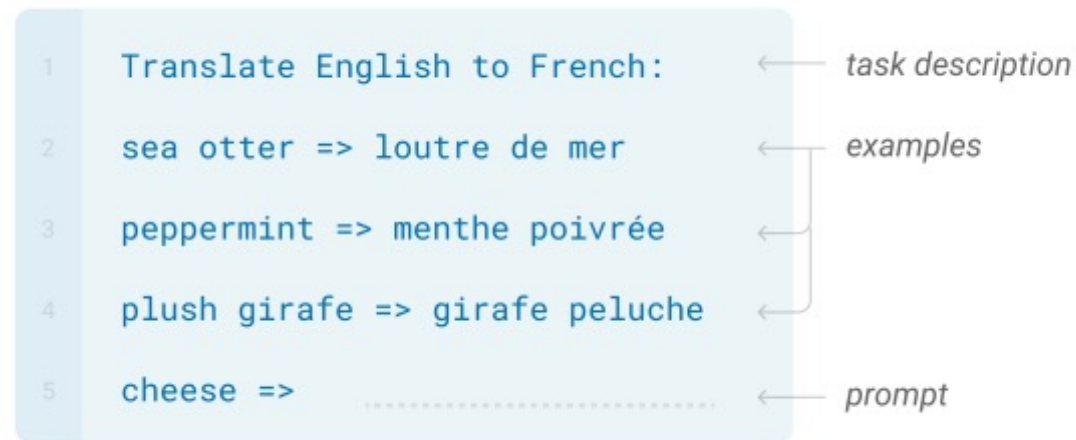
Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

# Nori et al. 2023

- 汎用LLMで特定の専門家の能力を引き出すために開発されたプロンプト戦略
- In-Context Learning (ICL) ・ Chain of Thought (CoT) ・ Ensembling (複数の出力結果から結論を得る) などプロンプトを工夫することで、汎用LLMがそのドメインに特化したLLMの性能を上回ることが出来ることを示した研究



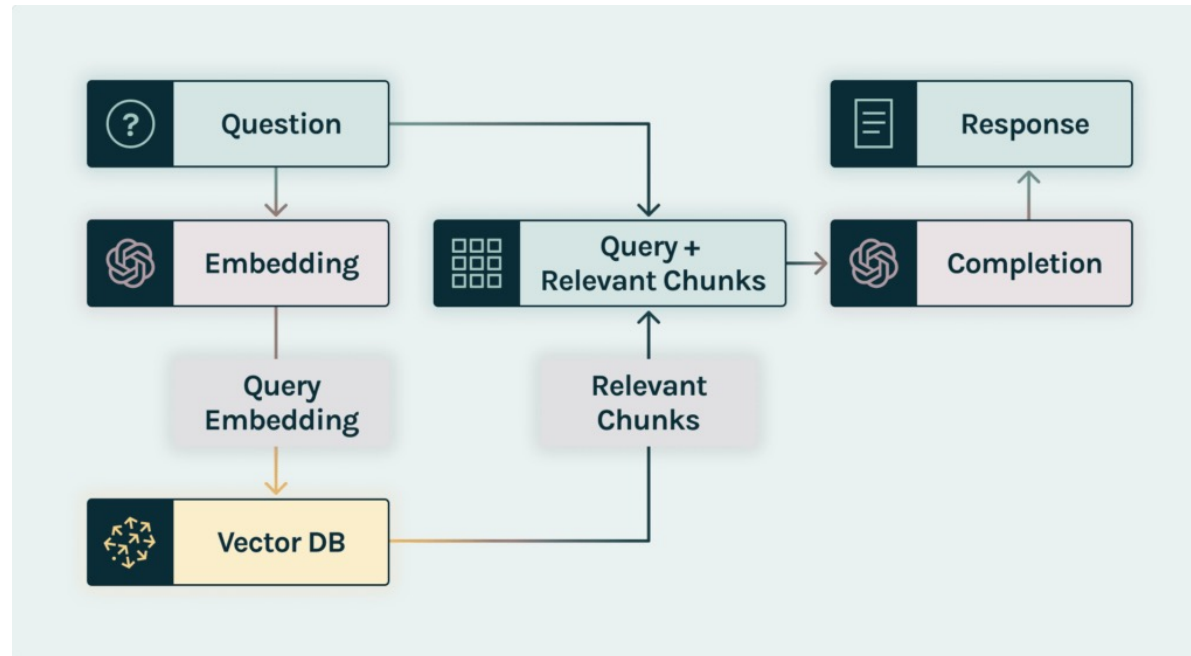
Brown et al. 2020

# Retrieval Augmented Generation (RAG)

- LLMの最新性・正確性・一貫性の課題を改善するアプローチの一つ
- LLMが持つ知識にすべて任せるのではなく、タスクに即して外部の知識を参照することで、タスクの性能改善（検索など）を目指す
- 単体のモデルに依存しない「Compound AI（複合型AI）」の手法として、実応用の分野でも注目されている

# Retrieval Augmented Generation (RAG)

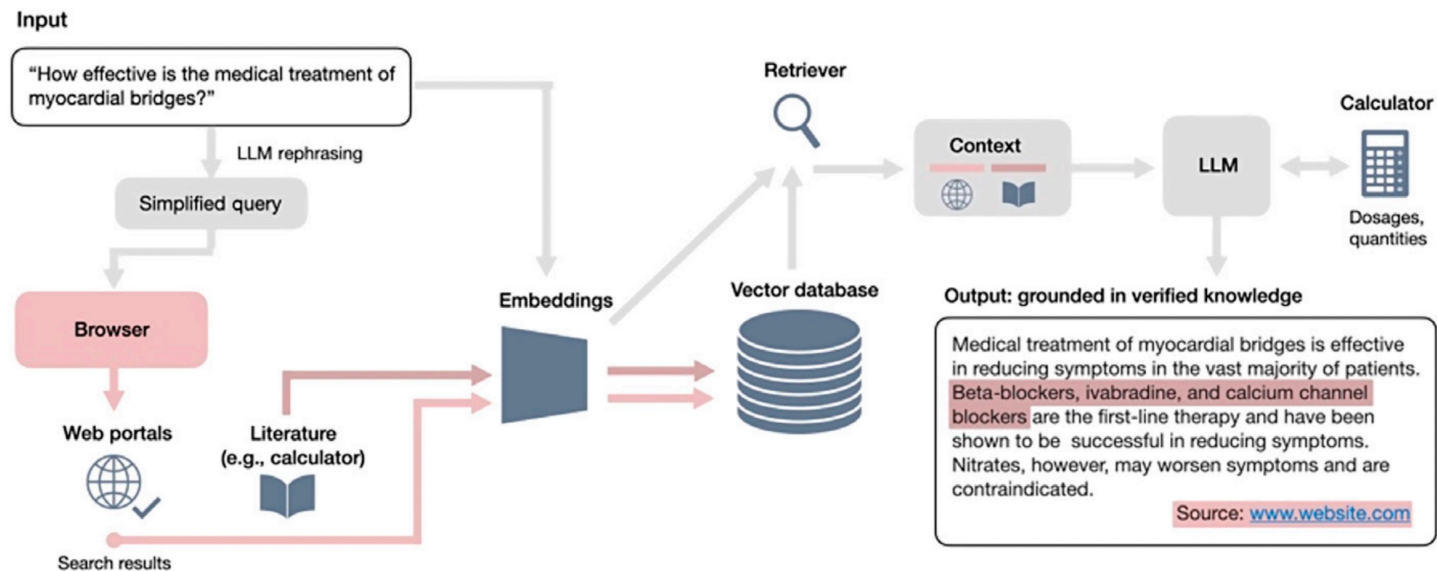
1. ユーザーのクエリーから、外部の知識を参照し、適切な情報を抽出する
2. その情報をプロンプトに注入し、LLMに応答を生成させる



<https://truera.com/ai-quality-education/generative-ai-rags/what-is-retrieval-augmented-generation-rag-for-llms/>

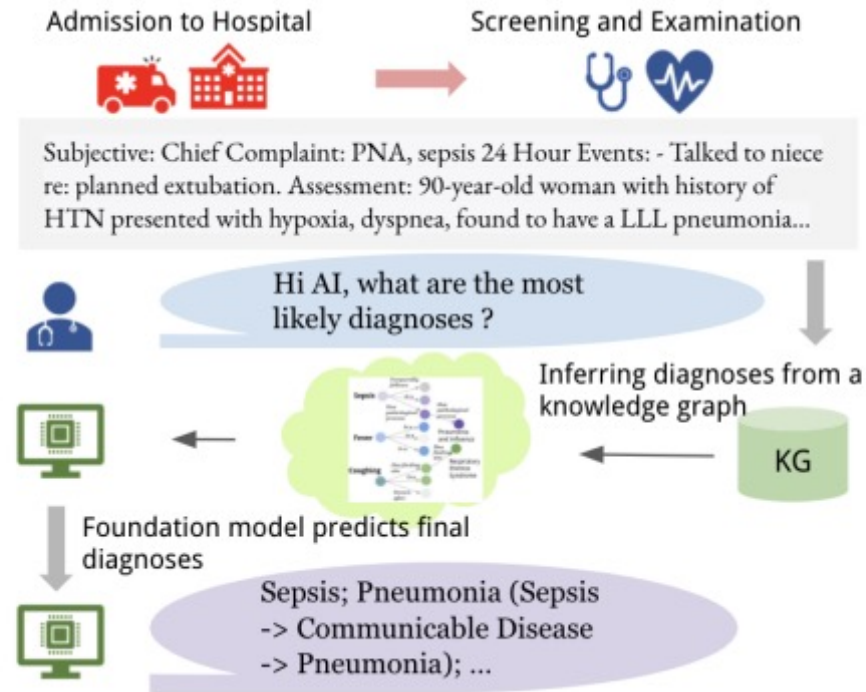
# Zakka Cyril et al. 2024

- RAGベースのLLMフレームワーク (Almanac) を提案
- PubMed、UpToDate、BMJ Best Practicesのウェブサイトを外部知識として使用
- 自身らでClinicalQAを開発し、一貫性などの複数の指標で有用性を示した



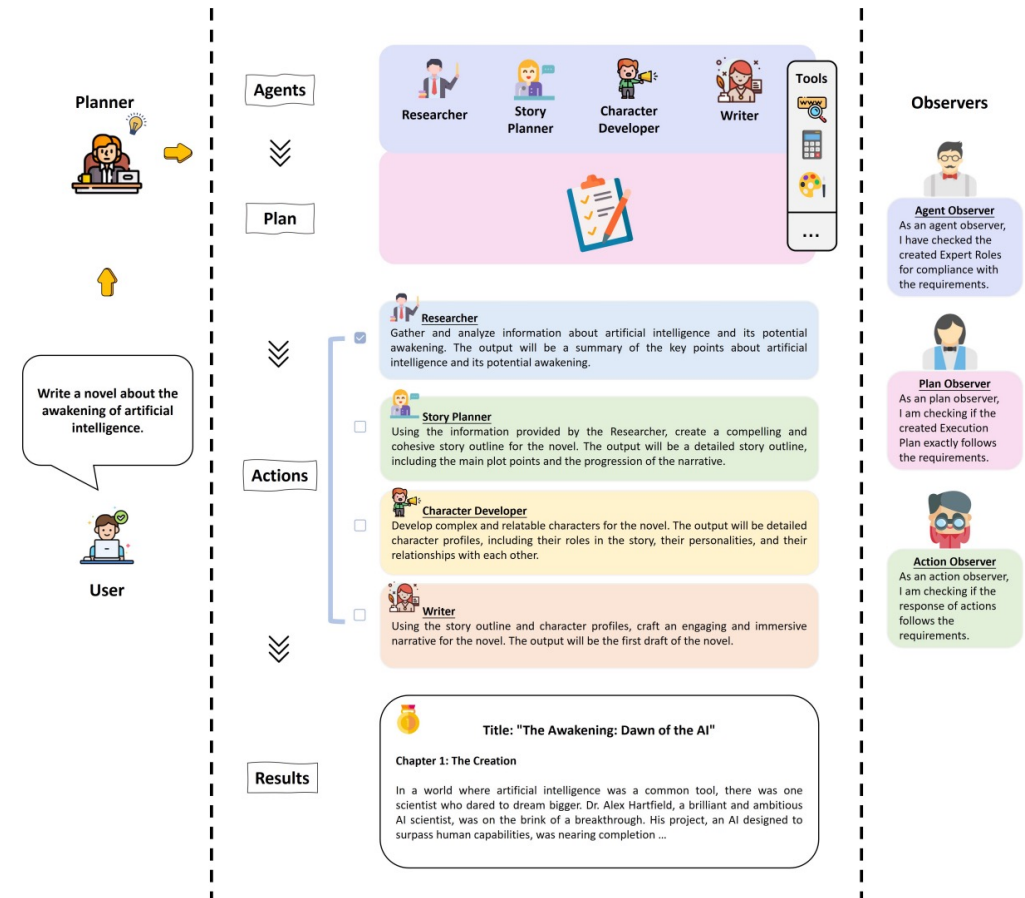
# Gao et al. 2023

- LLMの出力に医学分野の知識グラフであるUnified Medical Language System (UMLS) を活用して診断プロセスの合理性・正当性について追求し、性能改善を実現した研究



# LLM Agents

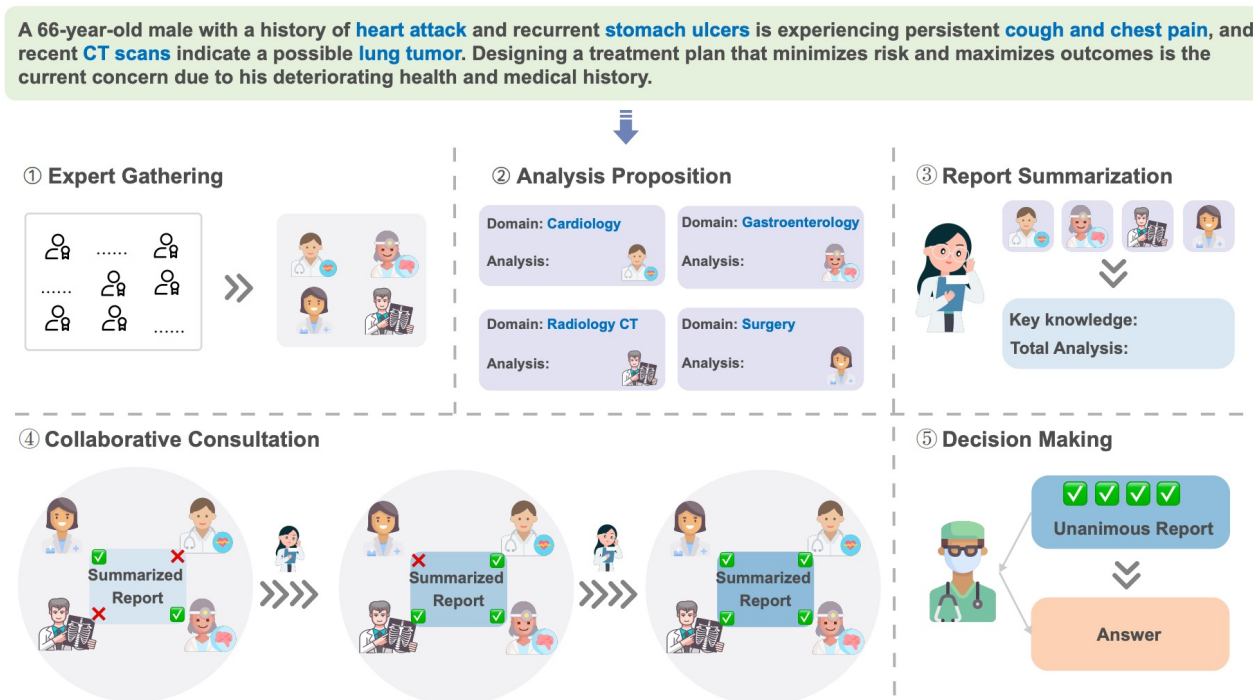
- LLMの発展により急速に盛り上がっている分野の一つ
- LLMを活用したマルチエージェントの実現により、ロールを与えられたエージェントがそれぞれの専門性を活かして、複雑なタスクの解決を目指す



Chen et al. 2023

# Tang et al. 2023

- 医療分野でのコラボレーション（MC）フレームワークを提案
- LLMエージェントを活用し、各ドメインが議論し、推論能力を高めることを示した研究





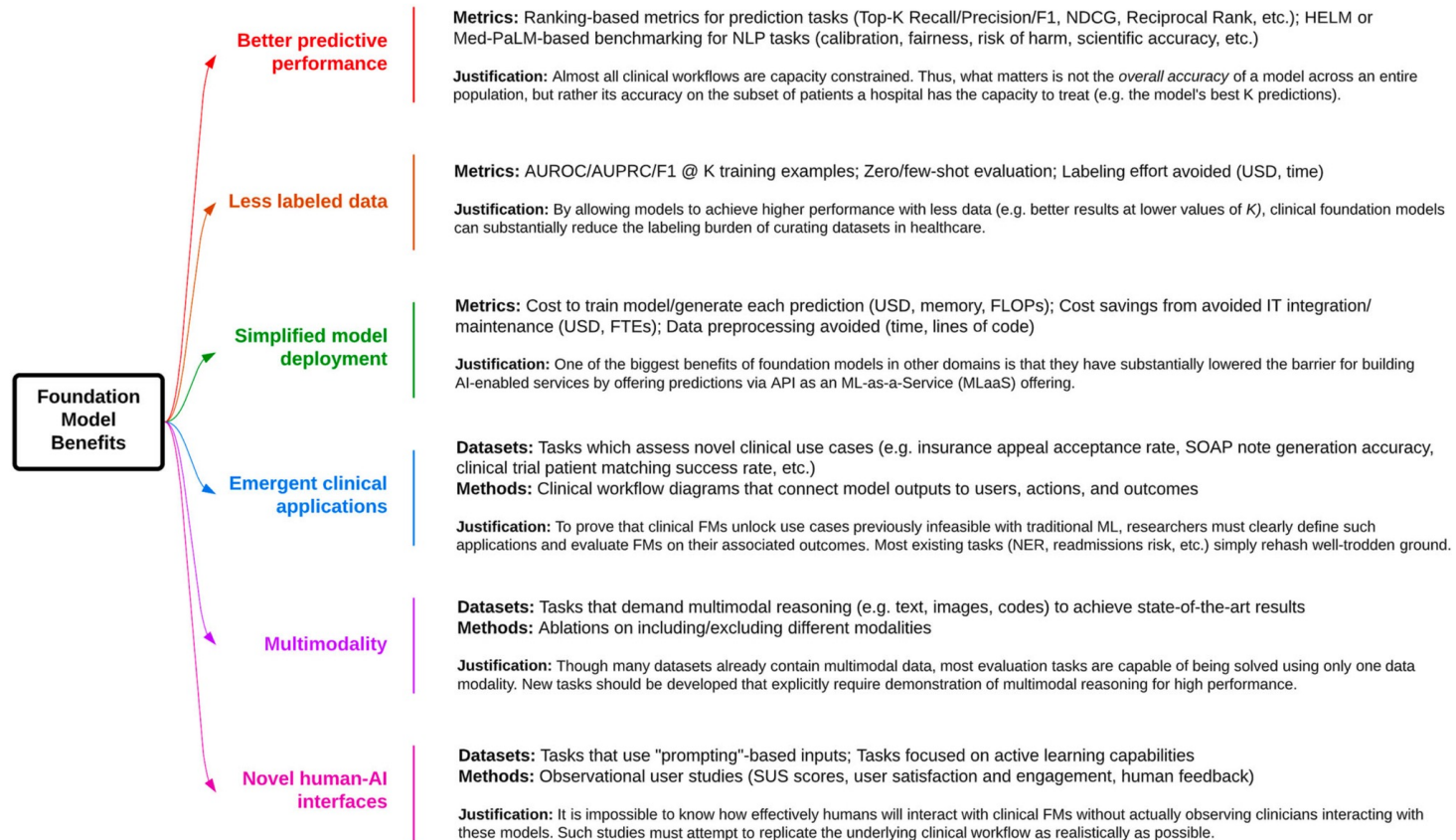
# Blagec et al. 2023

- 既存の医療領域のベンチマークと実臨床で求められるタスクのギャップを分析
- 臨床現場での大きな負担になっている日常的な文書化や患者データ管理のワークフローを評価するベンチマークはほとんど存在しておらず、臨床現場における負荷軽減を定量化するための新しいベンチマークの作成が必要となる

Clinical work activities	Task type	# Benchmarks <i>directly relevant</i> <i>(indirectly relevant)</i>
Write reports or evaluations	Administrative	3 (0)
Enter patient or treatment data into computers	Administrative	0 (0)
Develop treatment plans for patients or clients	Clinical care	0 (0)
Prepare official health documents or records	Administrative	0 (0)
Verify accuracy of patient information	Clinical care	0 (0)
Schedule patient procedures or appointments	Administrative	0 (0)
Process medical billing information	Administrative	0 (3)
Inform medical professionals regarding patient conditions and care	Clinical care	0 (0)
Prepare scientific or technical reports or presentations	Scientific	0 (0)
Verify that medical activities or operations meet standards	Clinical care	0 (0)
Review professional literature to maintain professional knowledge	Clinical care	4 (9)
Order medical supplies or equipment	Administrative	0 (0)
Process healthcare paperwork	Administrative	0 (0)
Gather medical information from patient histories	Clinical care	1 (9)
Process X-rays or other medical images	Clinical care	9 (24)

# Wornow et al. 2023

- 実臨床へのLLMの導入の意思決定のための実用的な指標を設計



# Thirunavukarasu et al. 2023

- 医療におけるLLM応用の実験的研究はほとんど行われていないため、革新的な使用例を実証し検証するための厳密な研究が強く求められている
- 具体的には、文書の質（有効な品質評価が必要）、作業効率、患者や医師の満足度などが考えられる
- 医療経済分析の観点からも、LLMアプリケーションの導入の費用対効果が高いことを立証することが求められる

# Minaee et al. 2024

## LLMの発展に向けた課題を以下の5つに整理

課題	内容
軽量化・効率化	蒸留 (Distillation) などにより、より軽量化し、コストを抑えられるモデルへ
アーキテクチャのパラダイムシフト	Transformer型からの発展 (Mambaなどの状態空間モデルへ)
マルチモーダル	テキスト・画像・音声などを組み合わせた利便性の高いモデルへ
拡張手法の発展	外部ツールと組み合わせたLLMの拡張・パーソナライズ
セキュリティ・倫理面への配慮	敵対的攻撃からの防御、倫理的面やバイアスへの対処

# 目次

- LLMの概要
- 医療分野におけるLLMの研究紹介
- 医療応用に向けたLLMへの期待と課題

## 推進に向けた期待と課題

1. 既存システムとのシームレスな連携
2. セキュリティ要件の議論
3. 個人情報等のプライバシー面の整備
4. 業務プロセスの改善
5. 情報抽出（データの集積）

# 技術イノベーションの浸透の難しさ

## **The Gap Between AI and Bedside** (Olaye and Seixas 2023)

スタートアップから生まれるイノベーションとその技術が臨床現場に統合される間に生じるギャップについて、以下の4つの課題を挙げている

1. 医療システムの調達プロセスに関する知識・理解
2. 厳しい規制や技術要件に関するコスト
3. 医療システムの調達プロセスに関する課題
4. 大企業と比較した、スタートアップ側の多面的な不利

# 大学病院とクリニックの環境の違い

## 大学病院

## クリニック

診療範囲

診療・手術・入院

診療がメイン

システム規模

大規模

小規模

調達プロセス

時間がかかる

短期な場合が多い

臨床研究（データの集積）

必要

行わない場合が多い

運用形態

オンプレミスが多い

クラウドが多い

**大規模病院への参入障壁の高さが課題**  
**→ 大規模病院の臨床現場に技術イノベーションが届きにくい**



## 既存システムとのシームレスな連携

- 大学病院など規模が大きくなるほど、システム連携が複雑になってくる
- 優れたソリューションであっても、既存システムとのインターフェースが不足しているため、導入できない事例もしばしば
- 各システムが連携しないことには、分離されたシステムを使用するユーザーの負担は大きく、現場には受け入れられにくい

# セキュリティ要件の議論

- 大学病院などでは、情報セキュリティの観点から病院情報ネットワークと呼ばれるクローズドなネットワーク内で多くのシステムが構成されている
- 昨今のセキュリティインシデントの事例から、ネットワーク境界のセキュリティにセンシティブな病院も多い
- 最近のLLMの多くはクラウドサービスを前提としていることが多いが、病院情報ネットワークが対策なしにオープンなネットワーク（インターネット）と接続されることは想定されず、そのギャップをいかに埋めるか（どう対策するか）の議論が不可欠になる

# 個人情報の取り扱い

- 病院などでLLMを利活用する場合、個人情報の扱いは避けられない
- 学会などもLLMへの関心は非常に高いが、個人情報の扱いに関するコンセンサスが得られておらず、利活用には慎重な姿勢を取っている
- 利用において、患者の同意、もしくは厳密な匿名化が必要とする場合、利用の障壁や匿名化による性能の影響などがあり、推進が難しくなる
- LLM推進のためには、現状の法規制・ガイドラインを正しく解釈し、どこまでが問題ないのか、どうしたら利用できるのかを議論していく必要がある

# 個人情報の取り扱い

- プロンプトに個人情報を含む場合、個人情報取扱事業者（病院など）がLLMサービス提供事業者に個人情報を「提供」（第三者提供）したと見なすか
- クラウド例外（Q&A7-53）基準を適用すると、サービス提供事業者が個人情報を取り扱わない場合、個人情報の第三者提供には該当しないとされる

クラウドサービスの利用が、本人の同意が必要な第三者提供に該当するかどうかは、クラウドサービスを提供する事業者において**個人データを取り扱うこととなっているかどうか**が判断の基準となります。当該クラウドサービス提供事業者が、当該個人データを取り扱わないこととなっている場合には、当該個人情報取扱事業者は個人データを提供したことにはならないため、**「本人の同意」を得る必要はありません。**

[個人情報取扱事業者の個人データに関する外部サービスの取り扱い（Q&A7-53）](#)

# 個人情報の取り扱い

- ただし、入力した個人情報がクラウド上に保存され、サービス提供事業者らがアクセスする可能性がある場合、前述のクラウド例外に該当しない
- サービス提供事業者がその情報をサービス改善に使用する可能性がある場合、個人情報の漏洩のリスクもあり得るため、より厳格な注意が必要とされる

個人情報取扱事業者が、あらかじめ本人の同意を得ることなく生成 AI サービスに個人データを含むプロンプトを入力し、当該個人データが当該プロンプトに対する**応答結果の出力以外の目的**で取り扱われる場合、当該個人情報取扱事業者は個人情報保護法の規定に違反することとなる可能性がある。

[生成 AI サービスの利用に関する注意喚起等について](#)

# 個人情報の取り扱い

- サービス提供事業者が個人情報にアクセスしない場合、個人情報の第三者提供にあらず、法律・ガイドラインに抵触せずに利用できる可能性が高い
- サービス提供事業者が個人情報にアクセスする可能性がある場合、契約条項／患者同意など個別のケースに応じて検討するのが望ましい
- サービス提供事業者がモデルの改善（学習）に利用する可能性がある場合、情報漏洩のリスクなどを鑑みて慎重に取り扱う必要がある

		改善（学習）に利用する	
		No	Yes
個人情報にアクセスする	No	●	
	Yes	●	●

個人情報・ガイドラインをもとにした解釈（あくまで個人の解釈）

## モデルの軽量化・効率化

- ネットワーク環境／個人情報課題を鑑みると、オンプレ上で動作する軽量モデルのほうが導入しやすい
- パラメータ数≒性能の世界において、実現は容易ではないが、現在も軽量化に向けて様々なアーキテクチャ・手法が提案されている
- 医療分野に特化したLLMの場合、一般ドメインを広く網羅する必要はなく、その点では軽量化の余地もあるのかもしれない

## 業務プロセスの改善

- 退院時サマリなどの文書作成・要約などによる業務支援への期待
- LLMに適したタスクされるが、実際の診療記録のサマライズには課題も多い
- 患者に必要な情報は診療記録・検査記録・画像診断レポートなど様々なデータソースに散在しており、また形式も多様
- このため、プロンプトとして与えるのに必要な情報を人手で収集し、準備する必要がある（文書作成にはこのプロセスに多くの時間がかかる）
- また、医師が記載する診療記録にはコピペ・文として成立しないような記載（メモなど）も含まれており、それらを解釈し、適切な文書を作成できるかについても確認が必要



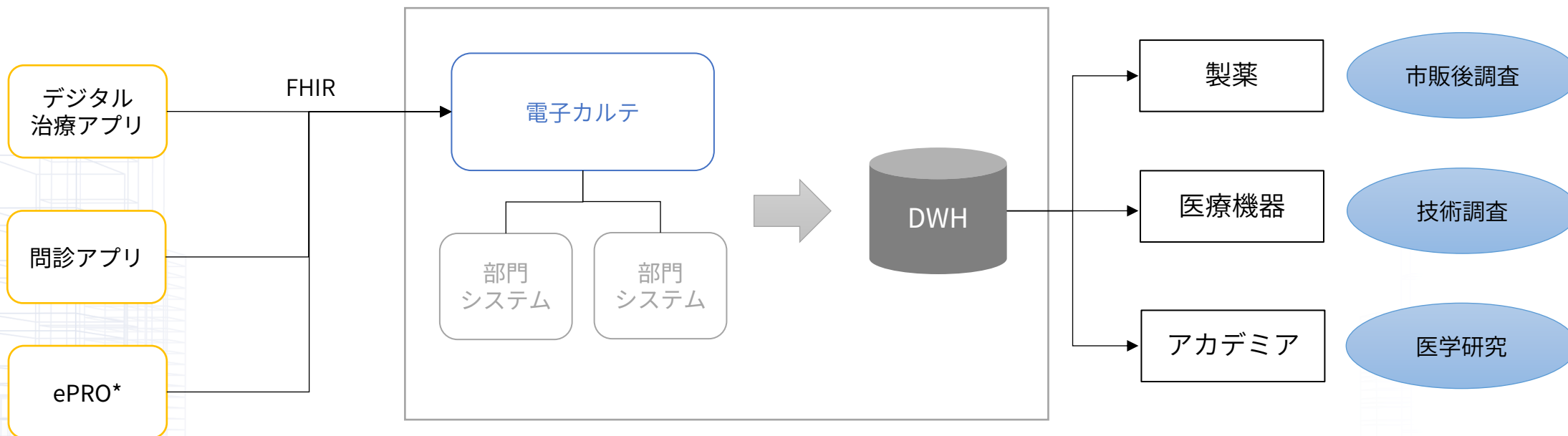
## 情報抽出（データの集積）

- 文書作成／診断支援以外にも情報抽出としてのLLMの可能性にも期待
- 前述の通り、患者に必要な情報を収集する難しさから、研究に必要なデータを揃えるのにも時間がかかり、またその範囲にも限界がある
- 作業なデータソースから、定型的な検索でできない（セマンティック検索に近い）情報抽出ができると、研究の推進にも貢献し得る
- また、自動化し、臨床プロセスに還元することで、モニタリング／病態検出などにも繋げることができる

# Big Picture

## デジタルヘルスによるシームレスな世界の実現に向けて

総合病院・大学病院など

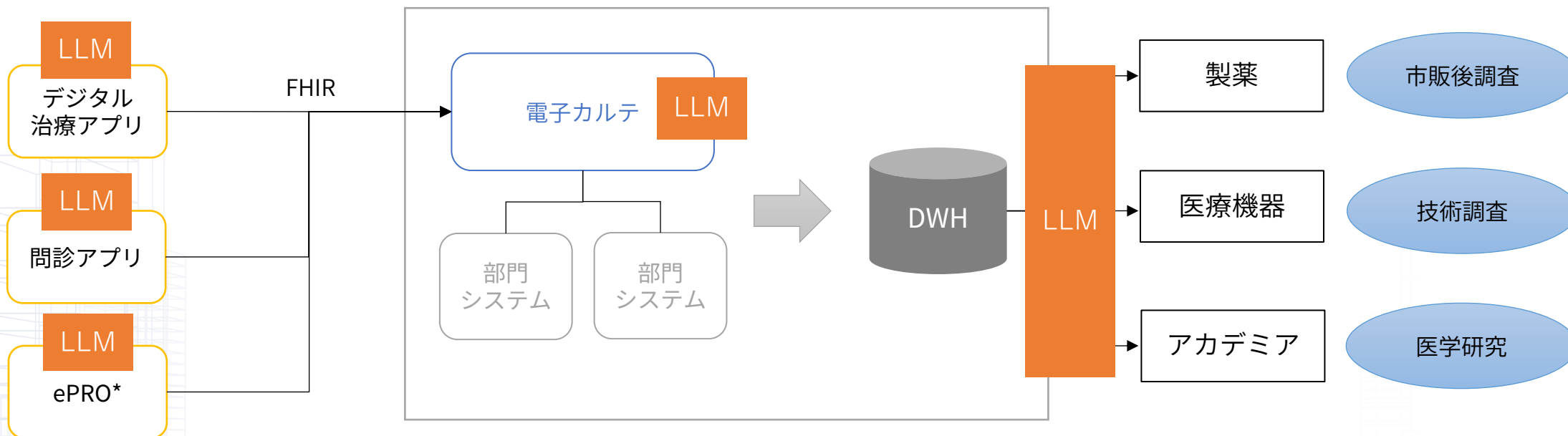


\*ePRO .. electronic Patient Reported Outcome

# Big Picture

## デジタルヘルスによるシームレスな世界の実現に向けて

総合病院・大学病院など



\*ePRO .. electronic Patient Reported Outcome

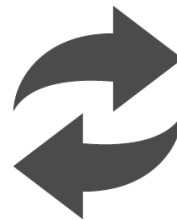
# 大阪大学医学部附属病院の取り組み

## データ・プラットフォームの構築

- デジタルヘルス／医療AI／研究にも多くの投資が必要
- 大学も診療報酬・研究費だけに頼らない収益の確保が求められる
- 病院データ・プラットフォームを構築し、第3の収益源につなげる

患者データなど（匿名化して提供）

大学病院



民間企業  
（製薬・医療機器等）

ライセンス費用・利用料など

# 大阪大学医学部附属病院の取り組み

## 阪大病院データバンク

- 人工知能等の新規技術の研究開発を推進するためには、民間企業・他の大学や研究機関と共に進める必要がある
- 一方、診療情報を用いた研究開発・商品化を企業等が行う場合、患者の同意がないと利用できない可能性があり、研究開発が速やかに行えないケースが生じている
- 医学研究及び医薬品・医療機器等の開発・商品化を国内外の研究機関や企業が共同もしくは単独で進めるために、患者から同意を得て取得した診療情報を蓄積し、利活用する



# 大阪大学医学部附属病院の取り組み

## 大阪臨床研究ネットワーク（OCR-net）

- 前向き研究のための臨床データ収集システム
- 後ろ向き研究のための共通データベース構築
- 保険データだけではなく、検体検査・画像データ・画像診断レポートなどを蓄積する仕組みを構築
- 製造販売後調査／医療機器開発などの利用に向けた仕組みを検討



19施設（2023年3月現在）が参加

# 参考文献

- Adams, Lisa C., Daniel Truhn, Felix Busch, Avan Kader, Stefan M. Niehues, Marcus R. Makowski, and Keno K. Bressem. 2023. "Leveraging GPT-4 for Post Hoc Transformation of Free-Text Radiology Reports into Structured Reporting: A Multilingual Feasibility Study." *Radiology* 307 (4): e230725.
- Arora, Anmol, and Ananya Arora. 2023. "The Promise of Large Language Models in Health Care." *The Lancet* 401 (10377): 641.
- Ayers, John W., Adam Poliak, Mark Dredze, Eric C. Leas, Zechariah Zhu, Jessica B. Kelley, Dennis J. Faix, et al. 2023. "Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum." *JAMA Internal Medicine* 183 (6): 589–96.
- Blagec, Kathrin, Jakob Kraiger, Wolfgang Frühwirt, and Matthias Samwald. 2023. "Benchmark Datasets Driving Artificial Intelligence Development Fail to Capture the Needs of Medical Professionals." *Journal of Biomedical Informatics* 137 (January): 104274.
- Chen, Guangyao, Siwei Dong, Yu Shu, Ge Zhang, Jaward Sesay, Börje F. Karlsson, Jie Fu, and Yemin Shi. 2023. "AutoAgents: A Framework for Automatic Agent Generation." *arXiv [cs.CV]*. arXiv. <http://arxiv.org/abs/2309.17288>.
- Clusmann, Jan, Fiona R. Kolbinger, Hannah Sophie Muti, Zunamys I. Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, et al. 2023. "The Future Landscape of Large Language Models in Medicine." *Communication & Medicine* 3 (1): 141.
- Gao, Yanjun, Ruizhe Li, John Caskey, Dmitriy Dligach, Timothy Miller, Matthew M. Churpek, and Majid Afshar. 2023. "Leveraging A Medical Knowledge Graph into Large Language Models for Diagnosis Prediction." *arXiv [cs.CL]*. arXiv. <http://arxiv.org/abs/2308.14321>.
- Jeblick, Katharina, Balthasar Schachtner, Jakob Dextl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, et al. 2023. "ChatGPT Makes Medicine Easy to Swallow: An Exploratory Case Study on Simplified Radiology Reports." *European Radiology*, October. <https://doi.org/10.1007/s00330-023-10213-1>.
- Kraljevic, Zeljko, Dan Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, et al. 2022. "Foresight -- Generative Pretrained Transformer (GPT) for Modelling of Patient Timelines Using EHRs." *arXiv [cs.CL]*. arXiv. <http://arxiv.org/abs/2212.08072>.

# 参考文献

Kung, Tiffany H., Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, et al. 2023. “Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models.” *PLOS Digital Health* 2 (2): e0000198.

Lee, Peter, Sebastien Bubeck, and Joseph Petro. 2023. “Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine.” *The New England Journal of Medicine* 388 (13): 1233–39.

Liu, Qianchu, Stephanie Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Maria Teodora Wetscherek, Robert Tinn, et al. 2023. “Exploring the Boundaries of GPT-4 in Radiology.” *arXiv [cs.CL]*. *arXiv*. <http://arxiv.org/abs/2310.14573>.

Minaee, Shervin, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. “Large Language Models: A Survey.” *arXiv [cs.CL]*. *arXiv*. <http://arxiv.org/abs/2402.06196>.

Nori, Harsha, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, et al. 2023. “Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine.” *arXiv [cs.CL]*. *arXiv*. <http://arxiv.org/abs/2311.16452>.

Patel, Sajan B., and Kyle Lam. 2023. “ChatGPT: The Future of Discharge Summaries?” *The Lancet. Digital Health* 5 (3): e107–8.

Qiu, Jianing, Lin Li, Jiankai Sun, Jiachuan Peng, Peilun Shi, Ruiyang Zhang, Yinzhaodong, et al. 2023. “Large AI Models in Health Informatics: Applications, Challenges, and the Future.” *IEEE Journal of Biomedical and Health Informatics* 27 (12): 6074–87.

Sarraju, Ashish, Dennis Bruemmer, Erik Van Iterson, Leslie Cho, Fatima Rodriguez, and Luke Laffin. 2023. “Appropriateness of Cardiovascular Disease Prevention Recommendations Obtained From a Popular Online Chat-Based Artificial Intelligence Model.” *JAMA: The Journal of the American Medical Association* 329 (10): 842–44.

Suchman, Kelly, Shashank Garg, and Arvind J. Trindade. 2023. “Chat Generative Pretrained Transformer Fails the Multiple-Choice American College of Gastroenterology Self-Assessment Test.” *The American Journal of Gastroenterology* 118 (12): 2280–82.



# 参考文献

Tu, Tao, Anil Palepu, Mike Schaeckermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, et al. 2024. “Towards Conversational Diagnostic AI.” arXiv [cs.AI]. arXiv. <http://arxiv.org/abs/2401.05654>.

Wang, Lei, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, et al. 2023. “A Survey on Large Language Model Based Autonomous Agents.” arXiv [cs.AI]. arXiv. <http://arxiv.org/abs/2308.11432>.

Wang, Sheng, Zihao Zhao, Xi Ouyang, Qian Wang, and Dinggang Shen. 2023. “ChatCAD: Interactive Computer-Aided Diagnosis on Medical Image Using Large Language Models.” arXiv [cs.CV]. arXiv. <http://arxiv.org/abs/2302.07257>.

Wornow, Michael, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A. Pfeffer, Jason Fries, and Nigam H. Shah. 2023. “The Shaky Foundations of Large Language Models and Foundation Models for Electronic Health Records.” NPJ Digital Medicine 6 (1): 135.

Yan, Michael, Giovanni G. Cerri, and Fabio Y. Moraes. 2023. “ChatGPT and Medicine: How AI Language Models Are Shaping the Future and Health Related Careers.” Nature Biotechnology 41 (11): 1657–58.

Zhou, Hongjian, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, et al. 2023. “A Survey of Large Language Models in Medicine: Principles, Applications, and Challenges.” arXiv [cs.CL]. arXiv. <http://arxiv.org/abs/2311.05112>.