

ChatGPT-5.2 と Google Gemini 3.0 の比較: 医療分野への応用可能性

ChatGPT5.2 版

概要 (Abstract)

OpenAI の ChatGPT-5.2 と Google の Gemini 3.0 は、最新世代の大規模言語モデル (LLM) として、医療分野での幅広い応用可能性を秘めている。本稿では、診断支援、医療文書要約、臨床意思決定支援、医療チャットボット、創薬支援、画像診断補助など医療業務全般への両モデルの適用可能性を評価する。評価軸として、正確性 (医学知識の一貫性・臨床的妥当性・

ファクトチェック能力)、安全性 (有害出力リスク・誤情報抑制・バイアス)、規則順守 (HIPAA・GDPR・日本の医療情報ガイドライン等への準拠、および医療用途での規制遵守) を設定した。各モデルのアーキテクチャや訓練方法、最新バージョンの性能仕様、医療分野での具体的ユースケースおよび第三者評価結果も含め、学术论文の形式で詳細に比較検討した。両モデルとも医療領域で極めて高い潜在能力を示しているが、ChatGPT-5.2 は高度な推論能力と信頼性で優れ、Gemini 3.0 はマルチモーダル処理や長大な文脈の処理に強みを持つことが明らかになった。それぞれのモデルは正確性・安全性・規制遵守の面で改良が重ねられており、医療現場での活用には人間の専門家による検証と組み合わせが必須であることも示唆された。

背景 (Background)

近年、大規模言語モデル (LLM) は医療分野に大きな変革をもたらす技術として注目を集めている[1]。OpenAI の ChatGPT シリーズ (GPT-4 以降) は医師国家試験や USMLE といった高度な医学試験で人間専門家に匹敵する成績を収め、医療相談や診断推論への応用可能性が示されてきた。一方、Google (Google DeepMind) は PaLM 2 を医療向けに調整した Med-PaLM 2 を開発し、さらにマルチモーダル AI である Gemini シリーズを展開している[2][3]。2025 年 11 月には、OpenAI が ChatGPT-5.1 (GPT-5 シリーズの最新版) をリリースし、直後の 12 月には改良版 ChatGPT-5.2 が発表された[4]。同時期に Google も Gemini 3.0 を公開し、AI モデルの性能競争はいわゆる「第二世代 AI 競争」とも呼べる様相を呈している[4][5]。本稿では、ChatGPT-5.2 と Gemini 3.0 という現行最先端モデルに焦点を当て、両者の医療分野への実務応用の可能性を評価する。特に、診断支援や文書要約など医療現場で想定されるユースケースにおける性能を評価軸 (正確性・安全性・規則順守) の観点から比較し、モデルのアーキテクチャや訓練プロセスの違い、第三者によるベンチマーク評価の結果を整理する。これにより、医療 AI として両モデルが持つ利点と課題を明らかにし、今後の医療現場での LLM 活用の指針を示すことを目的とする。

モデルのアーキテクチャと訓練 (Model Architecture and

Training) ChatGPT-5.2 の概要

ChatGPT-5.2 (GPT-5.2) は OpenAI が 2025 年 12 月にリリースした GPT-5 シリーズの最新モデルであり、同社の大規模言語モデルとして最も高度な性能を備える[4]。基本アーキテクチャは

Transformer ネットワークを基盤とする自己回帰型言語モデルで、前バージョン (GPT-5.1) から推論能力や長文コンテキスト理解がさらに強化されている[4]。OpenAI によれば、GPT-5.2 は一般知能 (汎用的な問題解決能力)、コーディング能力、長文コンテキストの理解で顕著な

改善を達成しており、複雑なマルチステップの作業処理にも優れる[4][6]。モデルのコンテキストウィンドウ長 (プロンプトに与えられるテキスト長) は GPT-4 の最大 32k トークンを大幅に超え、GPT-5.1 で 128k トークンに達していたとされる[7]。GPT-5.2 でも大規模コンテキストへの対応能力が高く、企業向け利用ではドキュメント丸ごとの解析も可能となっている (具体的なトークン長は非公開ながら 128k 相当かそれ以上と推測される)。さらに GPT-5.2 はマルチモーダルに対応し、テキストのみならず画像・音声・動画を入力として統合的に処理できる点で前世代と同様またはそれ以上に拡張されている[8]。このマルチモーダル能力により、例えば医学論文の図表解析や医療画像の説明にも対応可能である。

ChatGPT-5.2 の訓練には、インターネットから収集された大規模テキストコーパスに加え、人間フィードバックによる強化学習 (RLHF) が活用されている。OpenAI は GPT-4 で培った安全性チューニングの手法を継承し、GPT-5 シリーズでも大規模なレッドチーミング (専門家による攻撃的テスト) とアライメント調整を行っている[9][10]。GPT-5.2 はその集大成として、「これまでで最も信頼できるモデル」になるよう開発されている[11]。実際、OpenAI の内部テストでは GPT-5.2 は難解な質問に対しより少ない重大な誤りで回答し、幅広い分野で強力な性能を示したと報告されている[11]。また、GPT-5.2 では知識カットオフ (学習データの時点) が 2025 年 8 月まで更新されており、医学領域でも比較的新しい知見への対応力が向上している[12]。以上のように、ChatGPT-5.2 は最新アーキテクチャと高度な訓練手法により、医療分野の高度なニーズ (長文のカルテ解析や複雑な診断推論など) に応えうる基盤を備えている。

Google Gemini 3.0 の概要

Google Gemini 3.0 は Google (Google DeepMind) が 2025 年 11 月に発表した最新世代の AI モデルであり、Gemini シリーズの中でも最も高性能な Gemini 3.0 Pro が提供されている[13]。Gemini は当初からマルチモーダル処理とエージェント機能を念頭に設計されており、テキスト・画像・音声・動画といった複数の情報モーダルをネイティブに統合して扱える点が特徴である[13][8]。Gemini 3.0 では特に推論能力と長文コンテキスト処理が飛躍的に強化され、前バージョン (Gemini 2.5 Pro) よりも推論・コーディング・マルチモーダル分析の各ベンチマークで 50% 以上の性能向上を達成したとされる[14][15]。Gemini 3.0 のコンテキストウィンドウは業界でも前例のない 100 万トークン (約 104 万トークン) という桁外れの長さには達しており、非常に大規模な文書やコードベースを一度に解析できる[16][7]。例えば、膨大な電子カルテデータや数百ページに及ぶ治験プロトコル文書であっても、Gemini 3.0 は一度に入力として取り込み、全体を考慮した上で要約や解析が可能である。この長大な文脈処理能力は、医療のように情報量が多い領域では大きな利点となる。

Gemini 3.0 のアーキテクチャは、Transformer をベースにしつつ Google が培った様々な技術 (Mixture-of-Experts や強化学習によるツール使用等の要素) を組み込んでいると報じられている。モデルの具体的なパラメータ数は非公開だが、その性能から見て GPT-4/5 と同等かそれ以上の規模 (数千億から数兆パラメータ級) と推測される。Gemini 開発において特筆すべきは、マルチモーダル統合とエージェンシー (道具の利用能力) である[17][18]。

Gemini 1 では既に 画像とテキストのネイティブ処理や長文コンテキストの取り扱いに革新的な突破口を開き、Gemini 2 でエージェント的な振る舞い (ツールを呼び出して問題解決する能力) を基盤に据えた[19][20]。Gemini 3.0 ではその延長線上で、ツール使用や他システムとの対話を通じて複雑

なタスクを自律的に遂行できる「Deep Research」モードが導入された[17][18]。これは大量の情報源から知識を合成し、長時間にわたる推論チェーンを実行できるエージェント機能であり、例えば創薬研究における膨大な文献レビューや毒性評価などを自動化する用途が想定される[18][21]。Google は Gemini 3.0 をクラウドサービス (Vertex AI 等) として提供しており、開発者が API 経由でモデルにアクセスし自社アプリケーションに組み込めるよう設計している[17]。

Gemini 3.0 の訓練には、Google が保有する大規模データ (ウェブテキスト、学術論文、コード、画像・動画データなど) が用いられ、加えて人間フィードバックによる調整も OpenAI と同様に実施された。安全性向上のため、Google は「これまでで最も包括的な安全評価」を行い、Gemini 3.0 がエンタープライズ用途にも耐えうるよう厳格なテストとチューニングを実施したと述べている[22][23]。特に「Gemini 3 Pro は我々がこれまで構築した中で最も事実 に忠実 (factually accurate) なモデル」であり、複雑なタスクにおいて幻覚 (ハルシネーション) を最小化 するように訓練されているとされる[21]。また、1M トークンの長大コンテキストを扱うにあたって、新たな長距離依存関係の学習手法が用いられている可能性が高く、メモリ圧縮や階層型注意 機構などの研究成果が投入されていると推測される。以上の特徴により、Gemini 3.0 は長大 なマルチモーダル医療データを一括処理しつつ、AI エージェントとして外部ツールも駆使して 問題解決を図るという、次世代の医療 AI プラットフォームとしての潜在力を持つ。以下の表に、ChatGPT-5.2 と Gemini 3.0 の基本スペックを比較する。

項目	ChatGPT-5.2 (OpenAI)	(Google) Gemini 3.0 Pro
	モデル公開 2025 年 12 月[4]	2025 年 11 月[24]
アーキテクチャ	Transformer-based	マルチモーダル対応 テキスト、画像、音声、動画に対応[8]
自己回帰 LLM		
パラメータ規模	非公開 (推定数兆パラメータ級)	特徴的機能 Adaptive Reasoning (タスクに応じ思考深度調整)[25] 高度なコード生成・編集 高速レスポンス (Instant モード等)
コンテキストウィンドウ 入力:	~128,000 トークン	
出力:	 ~16,000 トークン[7]	

提供形態 ChatGPT (Web/アプリ) お づ対応 [13]

よび API 経由で提供

 (Instant・Thinking

・Pro の 3 モード) [26][27]

Transformer-based マ ルチモーダル

LLM 非公開 (推定数兆パラ メータ級) Vertex AI 経由の API 提 供

入力: ~1,048,576 トークン
出力: (Enterprise 向 け、各種ツール統合)

~65,536 トークン [7]

[17][28]

テキスト、画像、音声、動画にネイティ

表 1: ChatGPT-5.2 と Gemini 3.0 の主要スペック比較 [7][14]

医療分野での応用例と性能 (Applications in the Medical Domain)

次に、ChatGPT-5.2 と Gemini 3.0 の医療領域での具体的なユースケースへの適用可能性を、いくつかの主要なカテゴリ別に検討する。

診断支援と臨床意思決定支援 (Diagnostic Support & Clinical Decision Support)

診断支援とは、患者の症状・検査結果などから鑑別診断や確定診断を導くプロセスであり、臨床意思決定支援は診断後の治療方針やマネジメントの選択を支援することである。LLM は膨大な医学知識にアクセスできるため、これらの領域で医師を補助する AI ツールとして期待される。

ChatGPT-5.2 (および先行する GPT-4/5 モデル) は、臨床症例の解析と診断推論において極めて高い潜在能力を示している。例えば、Stanford 大学や UVA Health などが参加した研究では、GPT-4 搭載の ChatGPT が提示された複雑な臨床ケースについて、AI 単独で 92% の診断 正答率を達成したことが報告されている [29]。この精度は、同じケースで人間医師のみが行った場合 (正答率約 73.7%) を大きく上回り、医師 + ChatGPT 併用の場合 (76.3%) よりも高かつ

た [30][29]。研究者らは、この結果から「AI 単独でも診断に極めて有効なツールとなり得る」が、医師が効果的に AI を使いこなすためには更なるトレーニングが必要だと指摘している [31][32]。このように GPT-4 世代で既に高精度の診断推論が可能であったことから、知識量と推論力がさらに向上した ChatGPT-5.2 では、より難解な症例や希少疾患においても安定した診断支援が期待できる。ChatGPT-5.2 は OpenAI 独自のファクトチェック機能や補助ツール (ウェブ検索プラグイン等) と連携でき、診断に必要な追加情報を収集・検証しながら回答の妥当性を高めることが可能である。また、ChatGPT-5.2 は回答に不確実性がある場合にその 旨を示唆する傾向が強化されており (「最もあり得る診断は X ですが、さらなる検査 Y が必要です」など)、臨床意思決定支援においても慎重な提案を行うよう調整されている。OpenAI は GPT-5.2 を「困難な質問に対して最も質の高い解答を返すモデル」と位置付けており [11]、まさに医療のように正確性と慎重さが要求される領域において真価を発揮し得る。

一方、Google Gemini 3.0 も診断支援・意思決定支援への応用を強く意識して設計されている。Gemini は膨大な医学知識を学習しており、特に Google の医療特化モデルである Med-PaLM 2 や Med-Gemini の系譜を引き継ぐことで、高度な医学 QA 能力を備えている[33]。実際、Google が 2024 年に発表した医療版 Gemini (Med-Gemini) は、米国医師国家試験に相当する MedQA (USMLE 問題) において 91.1% の正答率を記録し、前世代の Med-PaLM 2 を 5% 上回り、GPT-4 も僅差で超えたと報告された[33] (図 1 参照)。これは Med-Gemini が医学領域に特化した追加訓練を受けた成果であり、Gemini 3.0 一般モデルにもその知見がフィードバックされていると考えられる。Gemini 3.0 は「最も事実に忠実なモデル」を謳っており、複雑な推論タスクでも不確かな情報を作り出す(幻覚する)リスクを低減するよう訓練されている[21]。例えば Google は、Gemini 3.0 を用いた研究用エージェント「Deep Research」を開発し、創薬の毒性評価やデューデリジェンスといった高度情報検索タスクで、モデルが安易に虚偽の情報を挿入しないよう工夫している[18][21]。さらに Gemini 3.0 の長大なコンテキスト処理能力は、過去の大量のカルテ情報や一連の検査結果を丸ごと踏まえた包括的な診断を可能にする。例えば、一人の患者の数年分の電子カルテ記録をすべて読み込ませ、その中から過去の症状推移・投薬歴・検査値の変化を踏まえた診断推論を行うことも現実的に可能である。実際、Gemini 3.0 の前身モデルを用いた研究では、長文の症例提示(数千語に及ぶ記録)に対しても一貫した推論が可能であることが示唆されている[34]。加えて、Gemini は外部ツールとの連携によって最新の医学ガイドラインや文献にアクセスし、エビデンスに基づく意思決定支援を強化できる。以上より、診断支援・意思決定支援において ChatGPT-5.2 は豊富な汎用医学知識と慎重な応答で優れ、Gemini 3.0 は長期にわたる患者情報の統合や最新情報の検索能力で強みを発揮すると言える。両者とも臨床現場での適用に際しては、医師による最終的な判断と検証を前提に補助的役割を果たすことになるが、その精度と知見の広さは医療従事者の意思決定プロセスを大いに支える可能性が高い。

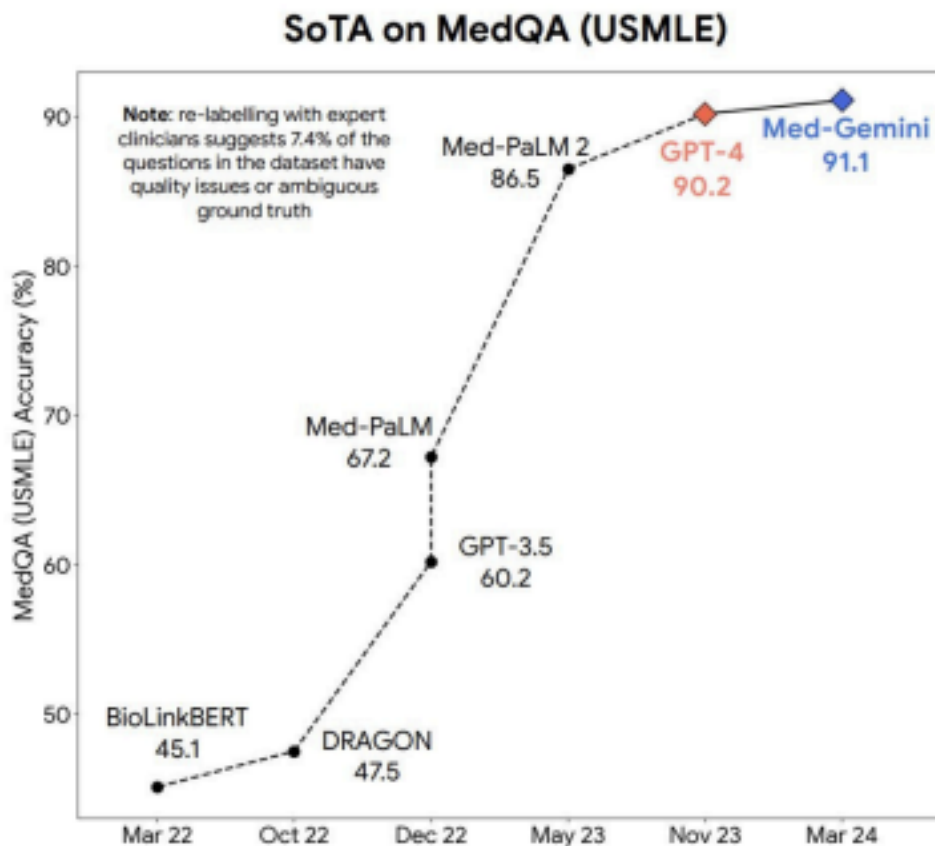


図 1: 医学質問応答ベンチマーク(MedQA/USMLE)における最新モデルの性能比較[33]。GPT-3.5 から GPT-4、Med-PaLM 2、Med-Gemini に至るまで正答率が大幅に向上している。2023 年時点で GPT-4 が約 90%の正答率を達成し、2024 年に発表された Med-Gemini では 91.1%に到達した(Med-PaLM 2 は 86.5%)。このように Gemini は医学領域で卓越した性能を示すが、GPT-4/5 も汎用モデルながら遜色ない精度を持つ。なお、質問データの一部には曖昧な正解や質の問題があることが注記されている。

医療文書要約 (Medical Document Summarization)

医療現場では、診療記録や検査レポート、退院サマリー、学术论文など大量の文書情報が日々生み出される。これらを迅速かつ的確に要約し要点を把握することは、医師・医療スタッフの負担軽減と医療の質向上につながる。LLM の高度な自然言語要約能力は、この領域で特に有用である。

ChatGPT-5.2 は、前世代モデルより長文要約性能が改善されており、長大な医療文書の自動要約に適している[35][36]。OpenAI の内部評価によれば、GPT-5.2 は数万字規模の文書をインプットしても文脈を保持したまま要約を生成でき、医療論文や教科書章などの要約において

情報の抜け漏れや誤解を減らすことに成功しているとされる[35][36]。特に「GPT-5.2 Thinking」モードでは長文入力に対する逐次的な要約と詳細な説明が得意であり、診療録か

ら患者の主要問題リストや経過の要約を抽出するなどのタスクに向いている[35]。実際、既に GPT-4 は米国の病院で電子カルテ要約に試験導入されており、医師の負担軽減に寄与している[37][38]。例えば HCA Healthcare では、医師と患者の会話記録から GPT を用いて自動でカルテ用のドラフトを作成する実験が行われており、医師はそれをレビューしてから電子カルテに転記するという運用で作業時間短縮が図られている[38]。GPT-4 搭載のこの仕組みでは、ハンズフリーの音声記録入力と組み合わせて対話からリアルタイムに診療メモを生成することに成功しており、2025 年にはより多くの病院で展開予定とされる[38][39]。ChatGPT-5.2 はこうした前例を踏まえ、要約結果の読みやすさ(要点を冒頭に明示する、箇条書きを適宜用いる等)もチューニングされている[40]。患者に説明するための要約(例えば難解な放射線科レポートを平易な言葉で要約)も可能であり、臨床コミュニケーション支援にも役立つと考えられる。

Gemini 3.0 も、大規模コンテキスト内での要約に卓越した能力を有する。Gemini 3.0 は 100 万トークンという膨大なテキストを扱えるため、患者の全入院経過記録や数十報の関連論文をすべて含めて入力し、それらを横断的に要約・統合することも潜在的に可能である[16][7]。例えば、一人の患者に関するプライマリケアから専門医までの記録を統合して「この患者の主要問題は何か」を抽出したり、特定疾患に関する最新エビデンスを多数の論文からまとめたりすることができる。MEDITECH 社(米国の EHR ベンダー)は、Google の LLM 技術を自社の電子カルテシステムに組み込んで患者記録の検索と要約を高度化する取り組みを進めており、Med-PaLM 2 やその発展系が用いられている[41][42]。具体的には、複数のシステムに散在するデータを統合し、患者の全履歴を縦断的に要約表示する機能や、退院サマリー・看護サマリーといった定型文書を自動生成する機能の開発が報告されている[41][42]。Gemini 3.0 はこうした用途において、より長い入力と多様なモーダルデータ(例えばカルテ内の波形や画像所見も含めた要約)に対応できる強みがある。さらに、Gemini は専門用語と一般語の言い換えにも長けており、対象読者に応じた要約を生成できる。例えば医療者向けには正確な専門用語で、患者向けには易しい表現で、といった調整を行うこともプロンプト次第で可能である。

以上のように、医療文書要約の領域では、ChatGPT-5.2 は既に実用化が進む臨床ノート要約で評価されており、Gemini 3.0 は桁違いの長文統合能力で将来的な大規模記録要約に適性を示す。それぞれのモデルは要約結果の質(正確性、一貫性、平易さ)に配慮したチューニングがなされており、今後は医療データ固有のフォーマット(処方箋、検査値一覧など)への対応も含めた洗練が期待される。

医療チャットボット (Medical Chatbot for Patient Interaction)

患者や一般市民からの医療相談に応じる医療チャットボットも、LLM の有望な応用先である。症状に関する助言、健康管理のアドバイス、服薬指導の補助、メンタルヘルス対話など、幅広い用途が考えられる。ただし、これらの用途では安全性と倫理が特に重要であり、誤情報や不適切な対応が直接人命に関わるリスクがあるため慎重な実装が求められる。

ChatGPT-5.2 は、自然な対話能力と幅広い知識により、人間らしい応答と共感的なやり取りを実現できる。GPT-4 ですでに患者との対話で思いやりのあるコミュニケーションが可

能であることが示されており[43]、GPT-5.2では会話のトーン調整機能(例えば口調をよりフォーマル/カジュアルにする等)も改善されている[40]。実例として、ある研究ではGPT-4を用いた対話エージェントが患者の訴えを傾聴し、適切なタイミングで共感を示しつつ医学的助言を提供することができたと報告されている[43][44]。ChatGPT-5.2ではこの傾向がさらに強化され、精神科領域でのカウンセリング的会話やセルフケア指導にも一定の有用性を持つと考えられる。ただしOpenAI自身も、ChatGPTを医療アドバイスに用いる際には必ず医療専門家の判断を仰ぐことを推奨している。GPTモデルは訓練データのバイアスや知識の限界から、不正確な情報をさも確からしく述べてしまう危険(幻覚)があり[45]、医療相談において誤った安心や不安を与えるリスクがあるためだ。実際、GPT-4でも稀に医学的に誤った指示(例えば禁忌薬の提案や緊急性の低評価)を出すことが指摘されている。GPT-5.2ではファクトチェック能力の向上や有害発言抑制のさらなる強化が図られているものの、「医師の代替」には依然なりえず、むしろ医療者の監督下で患者対応をサポートするツールとして位置づけられる。

Gemini 3.0もまた高度な対話システムとして、患者向けチャットボットの土台になりうる。GeminiはGoogle検索や医療知識グラフとの連携が比較的容易なため、患者が尋ねた症状に関して最新の信頼できる情報を取得し、回答に反映させることができると考えられる。例えば患者が「頭痛が続いていますが重大な病気でしょうか?」と質問した場合、Gemini搭載ボットは関連する一般的な原因(緊張型頭痛、片頭痛等)と重篤な徴候(髄膜炎の症状など)を挙げ、必要なら医療機関受診を勧めるといった適切なトリアージ対応が期待される。Googleは検索エンジンで長年にわたり健康情報パネルを表示してきた経験があり、その知見がGeminiにも組み込まれている可能性が高い。また、多言語対応にも優れるGeminiは、日本語を含む多数の言語で円滑に対話可能であり、地域の医療ガイドライン(例えば日本の診療ガイドライン)に沿った回答を生成することも訓練次第では可能である。事実、GoogleのMed-PaLM 2は専門医による評価で、回答の安全性・合意性の面で人間医師に匹敵するとされたケースもある[46]。もっとも、Med-PaLM 2の回答内容を米国医師と比較した研究では、一部で不正確な箇所や不十分な部分も認められており、LLMの実際の患者対応能力には限界があると指摘されている[47]。Gemini 3.0は幻覚抑制が強化されているとはいえ[21]、ゼロリスクには程遠く、患者に誤解を与える可能性は残存する。よって、安全な医療チャットボット運用には、(1)モデルが提供する情報に典拠や根拠を明示させる、(2)境界事例では回答を控え医師に相談するよう促す、といった安全策の組み込みが不可欠である。GoogleやOpenAIはモデルの出力に警告文を付す機能も提供しており、例えば「このアドバイスは一般的な情報であり、診断・治療の決定には医師の判断が必要です」といった免責メッセージを表示することが推奨される。

まとめると、医療チャットボット用途では、ChatGPT-5.2は自然な対話と共感能力で優れ、Gemini 3.0は情報統合力と多言語対応で強みを持つ。ただし双方とも、医学的正確性と安全性において完全ではなく、最終責任を負う人間医療者との協調が前提となる。現時点ではチャットボットが症状評価やセルフケア助言を行う際、LLMの提案は参考情報に留め、診断・治療の確定は必ず専門家が行うという運用が望ましい。

創薬支援 (Drug Discovery Support)

医薬品開発のプロセスにも、LLM は多角的に活用され始めている。創薬支援には、大きく分けて(1) 膨大な科学文献・データから知見を抽出して研究者をサポートする役割、(2) 化合物の設計やシミュレーションそのものに AI を用いる役割、の二面がある。LLM は主に前者の知識発見支援で力を発揮しつつあり、後者については分子生成専門のモデル(例えば蛋白質構造予測の AlphaFold など)との連携が期待される。

ChatGPT-5.2 は科学文献の理解と要約が得意であり、新薬ターゲットの探索や既存薬の転用可能性検討といったタスクで研究者を支援できる。例えば、新たな疾患 X に対する治療標的を見つける際、関連論文数十報を GPT-5.2 に要約させ、頻出する分子経路や有望なシグナル伝達経路を抽出する、といった使い方が考えられる。また、薬理データや臨床試験結果の表を読む

み込ませて結果の要点(有意差の有無や副作用プロファイル)をまとめさせることも可能である。GPT-4 は既に論文要約や実験計画の立案補助に利用され始めており、GPT-5.2 では知識カットオフが 2025 年まで延びたことで、最新の研究動向にも明るい[12]。さらに、ChatGPT プラグインや OpenAI API を通じ、ケミカルインフォマティクスデータベース(PubChem 等)に問い合わせで化合物情報を取得したり、Python ツールを呼び出してデータ解析を行ったりすることも可能である。そのため、研究者が ChatGPT を対話型のリサーチアシスタントとして用い、創薬に関する仮説出しやデータ整理を効率化する事例が増えている。ただし、LLM は生成した仮説の科学的妥当性を自動で検証する機能は持たないため、人間の専門的判断との併用が重要になる。

Google Gemini 3.0 は、Google Cloud 上で創薬支援 AI としての活用が想定されている。既に Bayer 社は Vertex AI および Med-PaLM 2 を使って創薬プロセスの効率化に取り組んでおり、Gemini への移行も視野に入れている[48][49]。具体的には、生成 AI によって研究者がデータ

にアクセス・検索・関連付けする作業を容易にし、大量の研究データから潜在的な関連を見出す試みがなされている[48]。例えば化合物 A に関する過去の特許や論文を一括分析し、その作用メカニズムに関する知識グラフを構築する、あるいは臨床試験のコミュニケーション文書を自動生成・多言語翻訳するといった用途である[48][50]。Gemini 3.0 は膨大な情報源の中から「針の糸を通す」ような関連性を見出すのに向いており、長大な文脈中で重要なポイントを抜き出す能力が創薬研究に資する。さらに、Google は高性能計算(TPU)を活用した量子化学計算やシミュレーションにも注力しており[51]、Gemini などの LLM による仮説生成と組み合わせることで、生成的 AI+計算科学のハイブリッドな創薬手法が模索されている。Gemini 3.0 自体も科学論文から知識を学んでいるため、新規分子標的に関する洞察や予測をテキスト生成という形で提供できる可能性がある。ただし、分子構造そのものの設計(デノボドラッグデザイン)には、LLM 単体では限界があり、専用の分子生成モデルとの連携が必要である点には留意が必要である。Google DeepMind は AlphaFold で示したように生命科学 AI に強みを持つが、Gemini とそれらを統合して使うフレームワークはまだ開発途上と言える。

総じて、創薬支援において ChatGPT-5.2 と Gemini 3.0 は主に情報処理と知識発見のツールとして有用である。ChatGPT-5.2 は対話形式で研究アイデアをブレインストーミングしたり

文献 調査を助けたりするのに適し、Gemini 3.0 は企業レベルで膨大なデータ資源を横断的に解析 するプラットフォームとして期待される。それぞれ直接に新薬を生み出すわけではないものの、研究者の創造的プロセスを強力に支援し、ひいては新薬開発の時間短縮やコスト削減に寄与することが期待されている[52][48]。

画像診断補助 (Image-based Diagnosis Assistance)

医療における画像診断(放射線画像、病理画像など)は、これまでディープラーニング(特に CNN)による専門モデルが主に研究・実用化されてきた分野である。近年、マルチモーダル LLM(GPT-4V や Gemini 等)が登場したことで、一つの統合モデルが画像とテキストの両面で 診断補助を行う可能性が出てきた。例えば、レントゲン写真や CT 画像を入力して所見を文章 化したり、画像所見と臨床文脈を総合して診断推論を行ったりすることが視野に入る。

ChatGPT-5.2(GPT-5 シリーズ)は画像入力機能を備えており、前世代の GPT-4V(ision)では既に医用画像に対する高精度の質問応答が確認されている。ある研究によれば、GPT-4V は医療画像質問応答(Medical VQA)のテストで 88%の正答率を記録し、同じテストでの人間専門医の精度(約 78%)を上回った[53]。一方、同テストで Gemini(Pro 版)は 61.45%に留まり、マルチモーダル推論において GPT-4V に劣後したとの報告がある[45]。この結果は 2024 年 前 半のものであり、Gemini 3.0 登場以前の比較ではあるが、少なくとも GPT-4 世代では医用画 像解析で OpenAI モデルが優位であったことを示唆している[45]。GPT-5.2 では視覚情 報の処

理能力がさらに強化されたと推定され、X 線画像や MRI を与えて異常所見を指摘させる、超音 波画像の所見をテキスト要約するといったタスクも高精度でこなせる可能性が高い。ただし、LLM による画像診断には注意点も多い。まず、学習データに含まれる医療画像の偏りや 限界 から、専門 AI と比べ見落としや誤検出が起こり得る。また LLM は画像内容を説明することは得 意でも、微細な病変の検知や定量的評価(例:腫瘍径の正確な測定)は専門モデルに譲る部 分がある。そのため、ChatGPT-5.2 を画像診断に使う際はあくまで補助的な「セカンドリー ダー」として位置づけ、専門医の読影を補完する役割で用いるのが現実的だろう。例えば、医 師が読影した後に ChatGPT にも所見をリストアップさせて照合し、見逃しを減らすといった利 用法が考えられる。

Gemini 3.0 はマルチモーダルモデルとして画像理解能力を備えるが、その詳細な性能は非公 開である。Gemini 開発陣は「現実世界の困難な条件下(話者の重複する音声、不鮮明な画 像)でも Gemini 3.0 は安定した性能を示した」と述べており[54]、画像認識においても堅牢性 を 重視していることが窺える[55]。Rakuten グループとのアルファテストでは、品質の低い文書 写 真から構造化データを抽出するタスクで Gemini 3.0 が従来比 50%以上の精度向上を示し たの エピソードが紹介されている[56]。これは医療画像とは異なる文脈だが、不鮮明な画像 から 情報を読み取る能力は示されている。医療応用に引きつけると、例えばスキャンが劣化 した古 いエコー画像や、低解像度の臨床写真からでもある程度の判断ができる可能性がある。また Gemini はエージェント機能を活かし、画像解析の専門ツールを自動で呼び出すこ

ともできる。例えば MRI 画像を外部のセグメンテーションモデルに渡して腫瘍領域を検出し、その結果を Gemini が受け取ってレポートを書く、というようなハイブリッド処理も将来的には可能となるだろう。さらに Gemini の 1M トークン文脈により、画像と共に膨大な関連テキスト(放射線検査の 依頼情報、患者の症状リスト等)を同時に考慮した総合判断が期待できる。これは人間の放射線科医が行う作業(臨床情報も踏まえた読影)に近づくものであり、純粋画像モデルには難しかった統合的診断支援が可能となる。

しかし現時点では、両モデルとも医療画像診断の専門 AI として認可されているわけではない。FDA や厚生労働省の承認プロセスを経た「医療機器 AI」ではなく、汎用モデルとして提供され

ているため、画像診断用途に使う場合には法的・倫理的な留意が必要となる(後述の規制順守の節で詳述)。総合的に見て、画像診断補助では ChatGPT-5.2 が現時点でややリードしている可能性があるものの、Gemini 3.0 も強力なマルチモーダル推論力で追随している。最終的には、これら LLM を既存の画像診断ワークフローにどう組み込むか(例えば読影レポートの

自動ドラフト作成や、症例カンファレンスでのディスカッション補助など)によって、その有用性が決まるだろう。

評価 (Evaluation of Key Criteria)

以上のユースケースを踏まえ、ここでは正確性・安全性・規則順守の 3 つの評価軸について、ChatGPT-5.2 と Google Gemini 3.0 を比較検討する。

正確性 (Accuracy and Clinical Validity)

正確性は、医療において最も重要なモデル評価基準である。モデルが提供する情報が医学的事実と合致し、臨床的に妥当で、一貫性のあるものかを評価する。

ChatGPT-5.2 の正確性: GPT-5.2 は、前述の通り一般医学知識や臨床推論問題において極めて高い正答率を示す。例えば医学試験問題集(MedQA)において GPT-4 が約 90% の正解率を達成したのに対し[33]、GPT-5.2 はさらに知識量と推論力が上積みされていると考えられる。OpenAI は GPT-5.2 について「様々な典型的ベンチマークで競合(特に Google)を凌駕する」と主張しており[57][58]、医学領域でも少なくとも定型化された課題において他モデル以上の正確性を発揮している可能性が高い。実際、TechCrunch が報じた独立ベンチマーク競争では、Google の新エージェントが独自の高度推論ベンチマークでトップとなったものの、OpenAI の ChatGPT-5 (Pro 版) が非常に僅差の 2 位につけ、ブラウザ操作タスクでは Google を上回ったという[59]。この事実は、GPT-5 世代が総合的なタスク性能で依然トップクラスであることを示唆する。また、GPT-5.2 は回答の一貫性も改善している。すなわち、同じ質問に対して状況設定を変えても矛盾の少ない回答を返し、内部知識の矛盾(ある場面で「A は正しい」と言い、別の場面で「A は誤り」と言うなど)を減らす努力が払われている。医学知識のアップデート面でも、2025 年 8 月までの知識を持つことで、新薬や新ガイドラインに関する回答の鮮度が増している[12]。もっとも、ChatGPT-5.2 も万能ではなく、特に専門領域の詳細では誤りを犯す可能性がある。例えば希少疾患や最新の試験結果など、トレーニングデータ

が十分でないトピックでは自信ありげに間違えた回答をするリスクがある。また、ファクトチェック能力は向上したとはいえ、自身の知識ソースを明示できるわけではないため、回答の裏付け確認は依然ユーザ側の責任となる。OpenAI は ChatGPT を医療助言の唯一の拠り所としないよう警告しており、これは正確性の限界を認めているとも言える。

Gemini 3.0 の正確性: Google は Gemini 3.0 Pro を「我々の中で最も事実ベースのモデル」とし[21]、特に長時間の複雑な推論タスクで幻覚を減らすよう訓練したと述べている[21]。このアプローチは、医療のように判断ステップの多いタスク(例: 複数の検査結果を経て最終診断に至る)で有効と考えられる。LLM は判断を下すステップが増えるほどどこかで間違える確率が上

がり、一つでも誤推論があると結果全体が崩壊するため[60][61]、Gemini のチューニングは医療応用に適した方向性と言える。実際、Google の医療特化版である Med-PaLM 2 は医療専門家からの厳しい評価にも耐える質を示し、臨床的妥当性で人間の回答に迫るとの報告もあった[46]。Med-Gemini の研究でも、医学 QA の正確性において GPT-4 や Med-PaLM 2 と同等以上である一方、視覚的質問応答や長文生成では既存モデルに及ばない面が指摘されている[45][62]。例えば、ある解析では Gemini Pro は診断正答率で Med-PaLM 2 や GPT-4 に劣らなかったが、画像を含む質問で精度低下が見られ、またモデルが自信過剰な誤答をする傾向が観察された[45][62]。この「自信過剰な誤答」は臨床的には危険であり、Gemini に残る課題と言える。また、Gemini 3.0 はトレーニングデータとして 2023 年までのデータを使っていると推測され、最新知識では GPT-5.2 に一步譲る可能性がある。もっとも、Google は 2025 年 11 月のリリース時に Gemini の性能を各種リーダーボードで強調しており、「Gemini 3 は AI モデル性能を測る人気指標の多くでリードしている」とも伝えられている[5]。具体的なベンチマーク名は明らかでないが、おそらく MMLU(学術知識テスト)や CodeBench(コーディング)、

HendrycksTest(専門知識テスト)などでトップクラスの成績を収めたのだろう。医療領域でも MMLU 医学カテゴリで高スコア、PubMedQA など高精度を出していると思われる。総合すると、Gemini 3.0 は正確性において ChatGPT-5.2 と同等レベルにあり、一部の特殊タスクで差が出る(画像では GPT 優位、最新知識では GPT 優位、長推論では Gemini 優位、といった補完的關係)と考えられる。したがって、医療応用時にはタスクに応じて両者の得意分野を踏まえ使い分けるのが理想となる。

安全性 (Safety: Harmful Outputs, Misinformation, Bias)

安全性は、モデルがユーザや患者に有害な影響を与えないか、誤情報やバイアスを拡散しないか、といった観点で評価される。特に医療では誤った指示による患者の健康被害や、偏見に満ちた対応による差別の助長などが懸念される。

ChatGPT-5.2 の安全性: OpenAI は ChatGPT の安全性向上に継続的に取り組んでおり、GPT-5.2 でも有害なコンテンツ出力の抑制が重視されている。具体的には、医療に関連する禁止事項(例えば自殺教唆や危険な治療法の提案など)に対しては応答を拒否する、あるいは「それにはお答えできません」といったスタイルで適切に回避するよう調整されている。GPT-4 のシステムカードでは、医療助言に関して不確実な場合に謝ったり警告したりする傾向

が組み込まれていたが、GPT-5 シリーズでも同様の安全策が踏襲されている[27]。さらに、OpenAI は GPT-5.2 を公開前に専門家によるレッドチームテストにかけており、医療シナリオも想定したテストを経ていると考えられる。例えば「子供が高熱ですがどうすれば？」という質問に対し、GPT-5.2 は適切に水分補給や医師受診を勧め、決して危険な民間療法を推奨しないような安全志向応答をするであろう。一方、幻覚問題については GPT-5.2 でも完全には解決されていない。知識の曖昧な領域では、それらしいが誤った情報を作り出してしまう危険は残る。そのため、医療用途ではモデルの出力を常に批判的に検証することが重要だ。GPT-5.2 は GPT-4 に比べ幻覚頻度が減ったとの指摘もあるが、それでもゼロではない。特に長い対話の中で一貫性を維持する過程や、ユーザが巧妙に禁止トピックへ誘導する場合に、予期せぬ不適切発言が生じるリスクがある。この点、OpenAI は「より穏やかで人当たりの良い会話スタイル」を GPT-5.1 以降で導入し、攻撃的・差別的発言の抑制も図っている[63]。結果、ユーザに敵意を向けたり、特定の属性への偏見を表現したりする確率は低く抑えられていると考えられる。総じて、ChatGPT-5.2 は安全性面で OpenAI が提示する最新基準を満たしており、明白に有害な出力は極力回避するようになっている。ただし「安全そうに見えて実は誤情報」というケース(例: もっともらしいが誤った医薬品用量の案内など)は依然として警戒が必要だ。

Gemini 3.0 の安全性: Google もまた AI の責任ある開発を掲げており、Gemini 3.0 では包括的な安全評価が実施されたと述べられる[22]。具体的には、Gemini 2 から 3 への進化でさらに多様な有害コンテンツフィルタを導入し、人種・性別などに関するバイアスや差別的表現を抑制する措置が強化されたと思われる。Gemini 3.0 はリリース前に数万人時に及ぶテストを課さ

れ、医学的アドバイスを含む様々なシナリオで適切に振る舞うよう微調整された可能性が高い。TechCrunch によれば、Gemini 3.0 は長時間自律動作するエージェントとしてもハルシネーションを最小化するように設計されている[21]。これは、エージェントが誤った行動方針を取らないよう、内部でのチェックポイントを設けていることを意味するかもしれない。医療応用では、Gemini エージェントが勝手に危険な診断や処方を実行しようとしないう制御されていることが望ましい。幻覚に関しては前述の通り、Gemini は依然いくらかの脆弱性を指摘されている[62]。特に自信過剰な幻覚は、安全性上 ChatGPT 以上に懸念かもしれない。モデルがあったかも事実であるかのように誤った情報を断言する場合、ユーザはそれを信じてしまう恐れがある。Gemini を医療で使う際は、こうした「一見もっともらしい誤り」を検知する仕組み(例えば、重要な数値やエビデンスには出典を提示させるなど)を組み込む必要があるだろう。また、Gemini は Google のポリシーに則り個人の健康情報についてセンシティブな扱いをすると思われる、過剰なプライバシー侵害に繋がる出力は抑制されることが考えられる。例えば特定の患者を想起させる実在データが入力された場合、それを外部へ漏らすような回答はしないよう、生成ルールが適用されているはずである(この点は規則順守の節でも触れる)。バイアス面については、現時点で Gemini 固有の医療バイアスは大きく報じられていない。GPT-4 では米国試験問題を訓練しているためか米国医療に偏った回答をする傾向があったが、Gemini/Med-PaLM は多国籍データを含み、よりグローバルな観点を持つとの指摘もある[64]。もっとも、訓練データの多くが英語情報である以上、非英語圏の医療知識や文化的感受性にギャップが残る可能性は否めない。

全体として、安全性に関して ChatGPT-5.2 と Gemini 3.0 は共に大幅な改善が図られており、医療現場で直ちに深刻な有害出力が飛び出すリスクはかなり低減されている。しかしゼロリスクではなく、人命を預かる医療領域では人間の二重チェックが不可欠という点は両モデルで共通する。むしろ安全な運用のためには、人間側がモデルの限界を理解し、適切なガードレール（許可される質問設定や出力制限）を設けることが重要である。

規則順守 (Regulatory Compliance and Privacy)

規則順守は、医療 AI を用いる際の法的・倫理的な要件への適合性である。ここでは特に患者情報の保護 (HIPAA、GDPR、日本の個人情報保護法等) や、医療機器としての承認、医療情報提供ガイドラインへの適合などの観点を扱う。

まず、データプライバシーの観点では、ChatGPT-5.2 も Gemini 3.0 も標準状態では医療機関が必要とする厳格なプライバシー基準を満たしていない。例えば米国の HIPAA (医療保険の相互運用性と責任に関する法) では、患者の個人健康情報 (PHI) を取り扱うサービスは事業提携契約 (BAA) を締結し HIPAA 準拠を謳う必要がある。しかし、OpenAI は現時点で ChatGPT 無料版・有料版 (Pro) について HIPAA 適合を保証しておらず、BAA も締結していない [65][66]。したがって、患者識別情報を含むデータをそのまま ChatGPT に入力することは HIPAA 違反となり得る [67]。OpenAI は企業向けの ChatGPT Enterprise プランや API 利用では入力データを学習に使用せず、必要に応じて BAA 締結も検討する姿勢を見せているが、医療現場で使うには契約上のクリアが必要である [68][66]。日本においても、患者の個人情報やカルテ情報をクラウド AI に入力することはガイドライン上グレーであり、多くの医療機関は慎重な立場をとっている。同様に、Google Gemini 3.0 も一般向け提供版 (Gemini Consumer) では HIPAA 非準拠とされる [69]。しかし Google の場合、医療機関向けには Google Cloud 経由で Gemini を提供し、BAA 締結下での利用を可能にする方針が示唆されている [70][71]。実際、Google Cloud は従来から医療向けサービスに HIPAA 対応を実装しており、Med-PaLM 2 も「HIPAA コンプライア

ンスをサポートするプレビュー版」が限定公開された [72]。Gemini 3.0 も Vertex AI 上での提供であり、適切なアクセス制御と監査ログを備えた形で使えば HIPAA や日本の個人情報保護指針に則って利用できると考えられる [70][71]。重要なのは、「クラウド上の LLM サービスを医療データに用いる際は、必ず暗号化やアクセス制限など安全措置を講じた上で、規制に準拠し

た契約形態で使う」ことである [71][66]。例えば、ある病院が ChatGPT API を電子カルテシステムに組み込むなら、OpenAI またはそのクラウド提供元 (Azure OpenAI など) と機密保持契約・BAA を交わし、やりとりされるデータは必要最小限かつ匿名化されたものに限定する、といった運用が必要となる。

次に、医療機器規制への適合について。米国 FDA や日本の PMDA は、診断や治療の意思決定に直接影響を与えるソフトウェア (SaMD: Software as a Medical Device) には承認を要求している。ChatGPT-5.2 や Gemini 3.0 は汎用 AI であり、このままでは医療機器として承認されていない。当然ながら、これらを医師の補助ツールとして用いること自体は違法ではないが、

それを患者への診断や治療の根拠として使うことには責任問題が伴う。例えば ChatGPT が提案した診断をもとに医師が治療しミスが起きた場合、その責任は誰が負うのか明確でない。現行の法制度では、AI は医療行為の主体にはなり得ず、最終的な責任は医師が負うことになる。ゆえに、医療従事者は LLM の提案を参考意見として扱い、裏付けを自ら確認することが求められる。また、患者への情報提供に LLM を使う際も、それが公式な医療説明と誤解されないよう注意すべきである。規制当局は今後、生成 AI を組み込んだ医療ソフトに関するガイドランスを整備するとみられる。実際、FDA は 2023 年に生成 AI の医療用途へのフレームワーク検討を開始している。日本でも、厚生労働省や医療情報学会が AI 診療支援システムのガイドライン策定に動いている。これらの規制が整うまでの過渡期においては、医療機関・開発者側で自主的に倫理指針を設けることが重要となる。ChatGPT や Gemini の提供元も、「医療用途での使用制限事項」を利用規約等で定めている可能性がある。例えば「診断を下す目的での直接使用は禁止」「出力は医療助言であって確定診断ではないと明示すること」といった条件だ。ユーザはこれを順守しなければならない。

最後に、ローカル規範への対応について。医療情報の取扱いには各国・各地域の規範がある。GDPR (EU 一般データ保護規則) は、健康に関する個人データの自動処理に厳格な要件を課している。LLM が欧州の患者データを扱う場合、匿名化してもなお個人特定が可能な情報が残っていないか注意しなければならない。日本では個人情報保護法に加え、病院等での電子カルテ情報の第三者提供に関するガイドラインがある。基本的に、患者同意無しにカルテ情報をインターネットサービスにアップロードすることは想定されておらず、LLM 利用時もそれに準じた患者同意の取得やデータ匿名化が必要となろう。また、日本の医療界には「AI を用いた診療はあくまで医師による最終確認を必須とする」旨の合意がある(例: 日本内科学会などの声明)。ChatGPT-5.2 や Gemini 3.0 のユーザは、こうした業界標準やエビデンスに沿った利用を心がけるべきである。

以上をまとめると、規則順守の面では ChatGPT-5.2 も Gemini 3.0 もまだ不十分な点があるが、それはモデル自体というより提供形態や利用方法に起因する。適切なエンタープライズ契約やプライバシー設定の下で使えば、両モデルとも HIPAA や日本の規制に準拠し得る [68][66]。ただし、現状では多くの場合、モデル出力をそのまま診療行為に結びつけることは規制上認められず、参考情報としての利用に留める必要がある。法規制やガイドラインが今後整備され、特定の用途に特化した LLM が医療機器承認を取得するようになれば、真の意味で「規則順守」と言える状況が訪れるだろう。それまでは、安全策を講じつつ慎重にこれら強力なモデルを活用していくことが求められる。

考察 (Discussion)

本稿で検討したように、ChatGPT-5.2 と Google Gemini 3.0 はいずれも医療分野での応用において極めて高い潜在能力を持つ。しかし、それぞれに得意分野と課題が存在し、単純な優劣では語れない複雑な状況が浮かび上がった。

性能面の総括: 正確性では両モデルとも既に専門医に匹敵する水準に達しつつある。GPT-5.2 は知識の幅広さと汎用性で強みを示し、Gemini 3.0 は深い推論と情報統合力で存在感を放つ。例えば診断推論では、GPT 系は豊富な事例知識の適用が上手く、Gemini 系は多面的

なデータ統合から新たな知見を引き出す力がある。要約では GPT が洗練された文章生成で優れ、Gemini はロングコンテキスト処理でアドバンテージを持つ。マルチモーダル処理では GPT-4V(5)が先行したが、Gemini も追随しつつあり、将来的には画像・音声を含む包括的な患者情報解析で両者が競う展開も考えられる。創薬支援など学術的タスクでは、どちらかと言えば Gemini は企業連携・計算資源を背景に大規模プロジェクト向き、GPT は個々の研究者が対話的に使うアシスタント向き、といった使い分けも見えてきた。このように、「ChatGPT vs Gemini」の構図は単なる性能競争に留まらず、ユースケース適合性の比較という観点が重要になる。医療業務には多様なニーズがあるため、最終的にはハイブリッド戦略(例えば、文章要約には ChatGPT を、データ検索には Gemini を使う、といった)が最も実利をもたらす可能性もある。

安全性と倫理の考察: 両モデルとも高性能化する一方で、依然として人間の専門知識と倫理観によるガバナンスが欠かせないことも明らかである。LLM は確率的生成モデルであり、与えられた入力に対しそれらしい出力を作るに過ぎない。そのため、たとえ正答率 90%を超えると

しても、残り数%のエラーが医療では致命的となり得る。モデルが出力する情報の信頼性を評価し、必要なら訂正する役割を医療従事者が担う必要性は、今後も不変だろう。むしろ、モデルが高度化するほど人間がうっかり盲信してしまう危険が増すため、クリティカルシンキングを持って AI と接する教育が医療者側にも求められる。また、患者との関係においても、AI が助言したからといってそれをそのまま伝えるのではなく、医師が解釈し咀嚼した上で説明責任を果たすことが重要だ。医療チャットボットに関しては、患者が AI を信用しすぎないように適切な境界設定(例えば「これは AI の仮の助言です」と明示する UI 設計)が必要になる。

規制と制度の展望: 現状、LLM の医療利用は制度的にはグレーな部分が多いが、各国で議論が進んでいる。米国では 2024 年以降、医療 AI の包括的な規制枠組みや認証制度が議会・FDA レベルで検討されている。EU も AI Act に医療分野の高リスク AI 分類が含まれており、LLM も適用対象となりうる。日本でも厚労省の研究班が AI 診療ガイドラインを策定中だ。こうした動きにより、ChatGPT や Gemini を直接または間接に組み込んだ医療システムが認証を取得して市場投入される可能性が出てくる。その際には、おそらく特定のタスク(例えば放射線科診断補助 AI)に LLM を活用した「狭い用途向け AI」として審査が行われるだろう。その場合、OpenAI や Google 自身が医療機器申請をするより、医療 IT 企業やスタートアップが両モデルを内包したシステムで申請する形が現実的かもしれない。いずれにせよ、技術の成熟に合わせた制度整備が追いつくことが、両モデルが本格的に医療現場へ普及する前提条件となる。

研究と第三者評価: ChatGPT-5.2 や Gemini 3.0 の医療応用については、今後も独立機関による評価研究が増えていくと思われる。既に引用したように、UVA や Stanford、Google 内部評価などいくつかの第三者評価が実施されているが、評価項目は限定的だったり環境が制御されていたりする。実地臨床でのプロスペクティブ試験、つまり AI を用いた診療と通常診療のア

ウトカム比較試験などはまだ報告が少ない。2024~2025 年には各国でそうした試験が始まる可能性があり、ChatGPT-5 系や Gemini 系が実際に診療成績を向上させるのかエビデン

スが蓄積されることが望まれる。あるいは逆に、想定外のエラーや人間側の過信による事故例が報告され、課題が浮き彫りになるかもしれない。それらポジティブ・ネガティブ双方の知見をフィードバックしていくことで、モデル開発者も医療者もより良い活用方法を模索できる。

競合他モデルとの関係: 本稿では ChatGPT と Gemini に絞ったが、医療向け LLM 領域には他にも Anthropic の Claude や各種オープンソースモデル (Llama 2 派生など) が存在する。例えば Claude 2 や Med-PaLM 2 は、対話の安全性や医学知識で独自の強みを持つとも言われる [73]。ただ現時点で、汎用性と性能のバランスで ChatGPT と Gemini が双璧をなしているのは間違いない。各モデルが互いに刺激しあう競争環境は、医療者にとっても選択肢が増える利点がある。将来的には、相互補完的に複数モデルを組み合わせた医療 AI システム (例えば、まず Gemini が情報集約し、その結果を ChatGPT がわかりやすく患者向けに説明文にする、といった) が出現するかもしれない。その意味で、ChatGPT vs Gemini は勝者が一方になるゼロサムではなく、協調進化の関係を築く可能性も示唆される。

結論 (Conclusion)

本稿では、2025 年末時点における OpenAI ChatGPT-5.2 と Google Gemini 3.0 を対象に、医療分野での応用可能性を多面的に比較した。正確性の面では、両モデルとも医学知識に基づく高度な回答生成が可能であり、定型的な医学試験問題では専門医と同等レベルの正答率を

示した。一方で長時間の推論や最新知識の反映といった点にわずかな差異がみられ、ChatGPT-5.2 は幅広い汎用知識と洗練された文章表現で優れ、Gemini 3.0 は長大な文脈統合と事実ベース推論で強みを持つことが分かった。[45][21] 安全性については、いずれも有害な出力や幻覚の抑制に注力しているが、依然として人間の監督が必要なレベルである。ChatGPT-5.2 は暖かみのある対話と誤情報低減策が講じられ [63]、Gemini 3.0 も幻覚を最小化すべく訓練されているが [21]、医療応用時にはモデル出力の検証と慎重な解釈が不可欠である。規則順守の観点では、両モデルとも現状では HIPAA や個人情報保護の要件を満たすには工夫が必要であり [66]、また医療機器として公式に認可されたものではないため、使用にあたっては契約面・法制度面で適切な対応が求められる。今後、法整備と運用ガイドラインが進めば状況は改善すると考えられる。

総合的に、ChatGPT-5.2 と Google Gemini 3.0 はいずれも医療 AI ツールとして有望であり、診断支援、臨床情報要約、患者対話、創薬研究など多岐にわたるユースケースで有用性を示すだろう。ChatGPT-5.2 はその高い汎用知能と安定した対話性能により、個々の医療専門職が

日常的に利用する「パーソナル AI アシスタント」として適している。一方、Gemini 3.0 は卓越したマルチモーダル統合力と長文処理能力により、医療機関や製薬企業が扱う大規模データを分析する「エンタープライズ AI プラットフォーム」として力を発揮すると期待される。それぞれのモデルの強みを踏まえ、適材適所で活用することで、医療現場の効率化と質の向上に寄与で

きる。重要なのは、人間の医療者と AI モデルが協調し、お互いの弱点を補完し合う形で医

療 サービスを提供することである。AI はあくまで意思決定を支援するツールであり、最終的な判断と責任は人間にある。この原則の下、ChatGPT-5.2 や Gemini 3.0 といった先端モデルを賢

く活用することで、私たちは医療という人類にとって極めて重要な領域において、より安全で効果的かつ人間味のあるケアを実現できるだろう。

参考文献 (References)

- OpenAI (2025) “OpenAI launches GPT-5.2 after 'code red' push to counter Google's Gemini 3” Reuters, Dec 11, 2025[4][5].
- OpenAI (2025) “ChatGPT Release Notes (Dec 11, 2025): GPT-5.2 launch” OpenAI Help Center[11][12].
- Google (2025) “A new era of intelligence with Gemini 3” Google AI Blog, Nov 18, 2025[14][15].
- Google Cloud (2025) “Gemini 3 is available for enterprise” Google Cloud Blog, Nov 19, 2025[22][55].
- TechCrunch (2025) Julie Bort, “Google launched its deepest AI research agent ... on the same day OpenAI dropped GPT-5.2” TechCrunch, Dec 11, 2025[18][21]. ● UVA Health (2024) Josh Barney, “Does ChatGPT Improve Doctors’ Diagnoses? Study Puts It to the Test” Medicine in Motion News, Nov 13, 2024[30][29]. ● Sergei Polevikov (2024) “Med-Gemini by Google: A Boon for Researchers, A Bane for Doctors” AI Health Uncut (Substack), Apr 30, 2024[33][47]. ● Pal et al. (2024) “Gemini Goes to Med School: Exploring the Capabilities of Multimodal LLMs on Medical Challenge Problems & Hallucinations” ClinicalNLP Workshop 2024[45][62].
- MoMo GPT (2025) “Gemini 3.0 と ChatGPT 5.1 を徹底比較” MoMo Co. Blog, Nov 26, 2025[16][7].
- HIPAA Vault (2025) “Is ChatGPT or Google Gemini HIPAA Compliant? A Complete Guide to HIPAA-Safe LLMs” (Blog)[68][66].
- Google (2023) Aashima Gupta & Greg Corrado, “How 3 healthcare organizations are using generative AI” Google Health Blog, Aug 29, 2023[48][38].

[1] [45] [53] [62] aclanthology.org

<https://aclanthology.org/2024.clinicalnlp-1.3.pdf>

[2] [3] [33] [34] [47] [64] Med-Gemini by Google: A Boon for Researchers, A Bane for Doctors

<https://sergeiai.substack.com/p/googles-med-gemini-im-excited-and>

[4] [5] [6] OpenAI launches GPT-5.2 after 'code red' push to counter Google's Gemini 3 | Reuters

<https://www.reuters.com/technology/openai-launches-gpt-52-ai-model-with-improved-capabilities-2025-12-11/>

[7] [8] [13] [16] [24] [25] Gemini 3.0 と ChatGPT 5.1 を徹底比較 | 性能・価格・使い分けを
完 全解説 - 株式会社 MoMo

<https://momo-gpt.com/column/gemini-3-0-chatgpt-5-1/>

[9] [10] Update to GPT-5 System Card: GPT-5.2 | OpenAI
<https://openai.com/index/gpt-5-system-card-update-gpt-5-2/>

[11] [12] [26] [27] [35] [36] [40] [63] ChatGPT — Release Notes | OpenAI Help
Center <https://help.openai.com/en/articles/6825453-chatgpt-release-notes> [14] [15]

[19] [20] Gemini 3: Introducing the latest Gemini AI model from Google
<https://blog.google/products/gemini/gemini-3/>

[17] [18] [21] [57] [58] [59] [60] [61] Google launched its deepest AI research agent yet
— on the same day OpenAI dropped GPT-5.2 | TechCrunch
<https://techcrunch.com/2025/12/11/google-launched-its-deepest-ai-research-agent-yet-on-the-same-day-openai-dropped-gpt-5-2/>

[22] [23] [28] [54] [55] [56] Gemini 3 is available for enterprise | Google Cloud Blog
<https://cloud.google.com/blog/products/ai-machine-learning/gemini-3-is-available-for-enterprise>

[29] [30] [31] [32] [43] [44] Does Chat GPT Improve Doctors' Diagnoses? Study Puts It
to the Test - Research - Medicine in Motion News
<https://news.med.virginia.edu/research/does-chat-gpt-improve-doctors-diagnoses-study-puts-it-to-the-test/>

[37] [38] [39] [41] [42] [48] [49] [50] [51] [52] [72] Google Cloud Next: Generative AI for
healthcare organizations
<https://blog.google/technology/health/cloud-next-generative-ai-health/>

[46] Toward expert-level medical question answering with large ... - NIH
<https://pmc.ncbi.nlm.nih.gov/articles/PMC11922739/>

[65] [66] [67] [68] [69] [70] [71] Using LLMs Under HIPAA: ChatGPT & Gemini
<https://www.hipaavault.com/resources/hipaa-compliant-hosting-insights/hipaa-compliant-llm-chatgpt-gemini/>

[73] Will GPT-5, Claude, Gemini Break Doctors' Monopoly On Medical ...
<https://www.forbes.com/sites/robertpearl/2025/09/02/will-gpt-5-claude-gemini-break-doctors-monopoly-on-medical-expertise/>