UNKNOWN LAB / RESEARCH REPORT 07

AI相談は、
専門職の代わりになれるのか

回答の上手さと、相談準備として安全に使えることを分けて考える

Daiki Morimoto / Unknown Lab / ORCID 0009-0009-6444-750X
対象限定レビュー / Version 1.0 / 公開日 2026年8月24日
DOI: 10.5281/zenodo.22065352 / 未査読

回答の上手さと、相談準備として安全に使えることを分けて考える

- Unknown Lab Research Report 07
- Version: 1.0
- 公開日: 2026年8月24日

- 著者: Daiki Morimoto / Unknown Lab
- ORCID: 0009-0009-6444-750X

- DOI: 10.5281/zenodo.22065352

- 種別: 対象限定レビュー。PRISMA準拠の系統的レビューではない。

30秒で読む

AIへ症状や生活の困りごとを書くと、すぐに、丁寧で、共感的に見える答えが返ってくる。それはすでに大きな価値である。

しかし、文章が上手いことと、専門職の代わりに診断、治療、緊急度、禁忌、本人の生活を判断できることは同じではない。

2026年の系統的レビューは、臨床LLM研究4,609件のうち、前向き無作為化試験が19件だったと報告した。公開質問への静的回答、試験問題、合成症例が多く、実患者の健康・生活結果、長期安全性、格差、責任まで検証した研究はまだ少ない。

一方、専門職へ相談する前に、経過、困りごと、質問、本人の希望を整理する用途には、限定的だが前向きな根拠がある。2,069人を対象にした多施設試験では、患者がAIで診療前レポートを作り、専門医が確認して通常診療を行う仕組みで、診察時間と対話の指標が改善した。

現時点で妥当なのは、AIを専門職の代役にすることではない。本人が相談を準備し、専門職が判断し、緊急時にはAIから離れ、誤りを訂正し、使わない経路を残すことである。

Care Compassも、この境界に置く。ただし、相談準備の効果も安全性も、Care Compass自体ではまだ検証されていない。

「文章が上手い」と「診療できる」の間

夜、家族の様子がいつもと違う。

救急車を呼ぶほどなのか。朝まで待ってよいのか。何を観察し、誰に、どう説明すればよいのか。医療機関へ電話する前に、AIへ尋ねたくなる場面はある。

AIは待たせない。言葉に詰まっても責めない。長い文章を短くし、質問を作り、何度でも説明する。忙しい専門職より共感的に見えることさえある。

問題は、その便利さが、AIに任せてよい範囲まで自動的に広げられやすいことである。

医師、作業療法士、看護師、心理職、ケアマネジャーなどの専門職は、文章だけを返しているのではない。観察し、身体や生活環境を評価し、追加情報を尋ね、過去の経過と照合し、不確実なまま責任を持って次の行動を選ぶ。本人や家族の希望と、制度や地域資源も調整する。

AIの文章を専門職と比べるなら、この仕事のどこを比べたのかを明確にしなければならない。

方法

2026年8月23日までに確認できた、患者・一般利用者向けLLM、相談前準備、患者質問への回答、緊急度判定、患者ポータル、診療前レポートの研究を対象にした。

一次資料、系統的レビュー、実患者を対象にした前向き試験、公的ガイダンスを優先し、WHO、厚生労働省、個人情報保護委員会、総務省消防庁、OpenAI公式資料と、Care Compassの公開コードを確認した。

研究は、次の5段階に分けて読んだ。

- 試験問題・合成症例
- 公開済み患者質問への静的回答
- 実患者データを使う後ろ向き・単群研究
- 実患者の前向き試験・無作為化試験
- 日常診療へ統合された運用と長期アウトカム

本稿は対象限定レビューであり、検索結果の全件スクリーニングや正式なrisk-of-bias評価は行っていない。特定製品の承認、安全性評価、医療または法的助言でもない。

4,609件の研究があっても、専門職代替とは言えない

Chenらは2026年、2022年1月から2025年9月までに発表された臨床LLM研究4,609件を系統的に整理した[1]。

- 実世界の患者データを使った研究: 1,048件
- 前向き無作為化試験: 19件
- シミュレーション: 1,857件
- 試験問題: 1,704件
- 人間と直接比較した1,046件で、LLMが人間を上回った研究: 33%
- 少なくとも4分の1の研究: 標本数30未満

この数字は、LLM研究が少ないという意味ではない。むしろ急速に増えている。

ただし、試験問題に正解すること、合成症例で専門家基準と一致すること、患者質問へ読みやすい文章を書くこと、実患者の相談行動や健康を安全に改善することは、それぞれ別の段階である。

高齢者向けの患者対話LLMに絞ったJohnsonらのliving systematic reviewでは、1,228件から採用されたのは9研究だった。参加高齢者数の中央値は12人で、7研究は支援的・社会情緒的な会話を扱っていた。臨床効果、費用効果、臨床ワークフローへ完全実装した状態を評価した研究はなかった[2]。

つまり、研究数の多さは、そのまま専門職代替の根拠量ではない。

共感的な回答は価値がある。ただし、安全性とは別に測る

よく引用されるAyersらの研究は、Redditのr/AskDocsから195件の患者質問と医師回答を抽出し、同じ質問へGPT-3.5が作った回答と比較した[3]。

3人の医療職が各回答を評価した585評価では、78.6%でチャットボット回答が好まれた。「良い・非常に良い」品質はチャットボット78.5%、医師22.1%。「共感的・非常に共感的」は45.1%対4.6%だった。

この差は無視できない。短く事務的な返信より、理由と次の行動を丁寧に書くことには価値がある。AIは専門職の文章作成を支援できるかもしれない。

同時に、チャットボット回答は中央値211語、医師回答は52語だった。評価者は正確性や捏造を独立項目で評価していない。公開フォーラムの単発回答であり、実際の診療関係、身体評価、患者転帰、受診の遅れは測っていない。

この研究が示したのは、特定条件で好まれやすい文章を作れることである。安全な診断・治療を任せられることではない。

実患者の試験が示したのは「診察前」である

2026年、TaoらはPreAという患者向けLLMを、中国の2医療機関で無作為化試験した[4]。

対象は24診療領域、111人の専門医、2,069人の患者・ケアパートナーだった。患者は専門診療の前にPreAと対話し、病歴、予備的な診断候補、検査提案を含む紹介レポートを作った。専門医はそれを確認してから、通常の対面診療を行った。

患者が補助者なしでPreAを使った群では、対面診察時間が平均3.14分だった。通常群は4.41分で、28.7%短かった。患者が評価したコミュニケーションの容易さは3.99対3.44、医師が評価したケア調整は3.69対1.73だった。

これは患者向けLLMの重要な前向き研究である。ただし、「補助者なし」は患者がAIを一人で操作したという意味であり、専門医が不要になったという意味ではない。

AIの後に専門医が診療した。主要評価は診察時間、医師評価のケア調整、患者評価の話しやすさである。診断エラー、長期安全性、健康・生活結果、費用効果を主要評価にした試験ではない。高負荷で診察時間の短い2病院の結果であり、在宅の自己相談や日本の医療へそのまま移せない。

それでも、この研究は一つの境界を具体化した。

AIが専門職の代わりに答えを確定するのではなく、本人の話を専門職へ渡せる形に整理する。

共創とfine-tuningは、同じ「チューニング」ではない

PreAは、患者、ケアパートナー、地域保健従事者、医師、看護師、管理者と共創された。低いヘルスリテラシーでも使える言葉、8-10往復の対話、専門医へ渡すレポート、失敗例への敵対的テストが調整された。

研究チームはRCTの前に、共創版と、同じモデルへ地域の診療会話515件を追加fine-tuningした版を、合成患者300例で比較した。共創版の方が、病歴聴取、診断、検査提案の評価が高かった。追加学習版は、地域の診療会話に含まれる省略や弱点まで再現した可能性があると報告された[4]。

この結果だけで「fine-tuningは悪い」とは言えない。実患者を共創対fine-tuningへ無作為化した比較ではなく、一つのデータセットと設定による合成実験である。

しかし、重要な警告は残る。

現場データを足すことと、現場の人がAIの目的・失敗・評価基準を変えられることは同じではない。

別のP&P Care試験では、中国11省の2,113人全員が同じAI相談を使い、その前にAIヘルスリテラシーのe-learningを受ける群と受けない群が比較された。e-learning群は、AI対話から評価された健康ニーズ認識が2.95対2.34で高かった[5]。

これも、AIが通常診療より有効だった試験ではない。AIを使う前の準備が、AIとの関わり方を変え得るという研究である。通常診療との無作為比較はなく、長期健康アウトカムもない。農村部では効果が小さく、デジタル格差は教育だけで消えなかった。

緊急度判定が、最も明確な境界になる

「念のため専門職へ相談してください」と書くAIは、一見安全に見える。

しかし、免責文があることと、緊急度が正しいことは違う。

Draelosらは、患者が尋ねる形のプライマリケア質問222件へ、4種類の公開LLMが答えた888回答を医師主導でred-team評価した。問題ありとされた回答はモデルにより21.6-43.2%、unsafeは5-13%だった[6]。これは実患者の有害事象率ではないが、もっともらしい回答の中に深刻な危害可能性が残ることを示す。

Reisらは227件の健康質問に対する908回答を調べ、97%が専門職への相談を促したと報告した。一方、質問の緊急度に応じた紹介表現はモデル間でばらついた[7]。

成人救急トリアージ15研究のメタ分析では、統合精度はChatGPT-3.5で0.51、GPT-4系で0.70、prompt最適化後で0.81だった。ただし研究間の異質性が高く、GPT-4の見かけ上の優位も頑健ではなかった[8]。

さらに、合成症例を使った安全監査では、危険な方向の誤りが見つかっている。

- 日本語の30症例、60ペルソナ、3,342出力の監査では、緊急red-flag症例の主要行動がほぼ全てunder-triageだった[9]。
- ChatGPT Healthの60症例、16条件、960回答のテストでは、gold-standard emergencyの52%をunder-triageした。利用者が受診を控えたい言葉を加えると、より低い緊急度へ寄るオッズが11.7だった[10]。

どちらも合成症例であり、一般利用者全体の失敗率ではない。それでも、同じ回答を再現できることが、安全性を保証しないと分かる。

一方、Pasliらの前向き多施設研究は、6,657人の救急患者で、主訴、併存症、バイタルを構造化し、病院別promptを用いた。GPT-4oと救急専門医基準の一致はkappa 0.833で、トリアージ担当者の0.782より高かった[11]。

この研究は可能性を示す。ただし、一般利用者が自宅から自由文で相談したのではない。病院内の構造化データと専門職のワークフローが介入の一部だった。複雑な高次医療の黄・赤区分では性能が下がった。

結論は、AIトリアージが常に危険でも、専門職より常に優れるでもない。

対象、入力、モデル版、prompt、基準、確認者、利用環境を含む仕組み全体を評価しなければならない。そして一般利用者向けでは、緊急時にAI会話を続けず、人間の窓口へ移る境界が必要である。

WHOも患者向けLMMについて、誤った・不完全な情報、プライバシー侵害、専門職との関係の劣化、ケアが医療システム外へ移る危険を挙げ、自治、安全、透明性、責任、公平性、持続可能性を統治原則にしている[20]。これは効果研究ではなく、設計と統治の最低条件を示す公的ガイダンスである。

相談準備には、AI以外から小さく確かな根拠がある

診察前に質問を整理する方法は、AI以前から研究されてきた。

Question Prompt List（QPL）は、患者が専門職へ聞きたい質問の候補をまとめたリストである。2015年のレビューは42研究、50介入を整理し、医師が支持し診察直前に渡すQPLは、質問数や専門職からの情報提供を増やす可能性を示した。ただし、知識想起、不安、満足、診察時間の結果は一貫しなかった[12]。

外科相談のRCTでは58人を解析し、QPL群で質問が24%多く、想起した情報項目が9%多かった[13]。

進行がんのQPLを扱った7 RCT、1,059人のメタ分析では、質問数と将来への期待の表明は増えた。しかし、健康状態、終末期アウトカム、不安、QOLの有意な改善は示されなかった[14]。

糖尿病患者1,276人のpragmatic cluster RCTでは、診察前に優先事項を1-2点提出する仕組みで、質問準備と治療選択肢を提示されたという報告は増えた。一方、12か月のHbA1c改善は通常群と差がなかった[15]。

ここには、AI相談の評価に使える重要な教訓がある。

相談準備は、質問、想起、話しやすさとして測れる。しかし、それだけで健康や生活が改善したとは言えない。

Care Compassが目指すのは、この相談準備に近い。AIが答えを確定するのではなく、本人が大切にしたい作業、困る場面、環境、家族、支援、道具、安全面を整理し、専門職へ渡すメモを作る。

ただし、既存のQPL研究を、Care Compassの効果証拠として流用してはいけない。Care Compass自体で、相談メモの正確性、訂正しやすさ、質問準備、専門職への持参、受診遅延、緊急離脱、個人情報入力、紙経路利用を検証する必要がある。

Care Compassは、何を軽減し、何を保証できないか

2026年8月23日時点の公開実装では、Care Compassは次を組み合わせている。

- 医療診断・緊急対応ではないと、開始前と利用中に表示する。
- 119、#7119、心理的危機の人間相談先を常時表示する。
- 一部の緊急語を規則で検出し、LLMへ送る前に会話を止める。
- AIの役割を相談準備、選択肢整理、専門職へ渡すメモに限定する。
- 個人が特定される情報を入力しないよう表示する。
- 画面上の履歴をブラウザのメモリ内で扱い、Unknown Labのアプリ用データベースへ保存しない。
- OpenAI APIへstore:falseを指定する。
- 紙のワークシート版を用意する。
- 利用者がリセットし、ページを閉じて停止できる。

これらは、用途誤認、一部の明示的危機、Unknown Lab側の永続保存、AIしか選べない状況を軽減する可能性がある。

しかし、緊急語の規則判定は、言い換え、誤字、婉曲表現、文脈を完全には扱えない。LLMが作る整理やメモも誤り得る。紙版があっても、アクセシビリティや相談行動が改善するとは限らない。

データについても限界がある。個人情報保護委員会は、一般利用者に対し、生成AIへ入力した個人情報が学習に利用されたり、正確または不正確な形で出力されたりするリスクを踏まえて判断するよう注意喚起している[16]。

OpenAI公式資料では、APIの標準的な不正利用監視ログに入力・出力が最長30日含まれる可能性がある。store:falseは、Responses APIのアプリケーション状態保存を抑える指定であり、承認済みZero Data Retentionと同じではない[17]。

Care Compassの現在の設計は、危険をゼロにするものではない。用途を狭くし、人間へ離脱し、データを減らし、代替経路を残すrisk-reduction designである。第三者評価はまだない。

日本では「一般情報」と「個別判断」の境界もある

厚生労働省のオンライン診療指針Q&Aは、一般的な医学情報や一般的な受診勧奨と、個別具体的な症状に基づく疾患可能性の提示・診断を区別する。後者は医学的判断を含み、遠隔健康医療相談としては実施できないと説明している[18]。

具体的なAIサービスへどの法令・指針がどう適用されるかは、提供主体、表示、行為、データ、医療機器該当性等で変わる。本稿は法的判断を行わない。

それでも、Care Compassを個別診断ではなく相談整理に限定し、緊急時は119や、実施地域では#7119へ離脱させる理由は明確になる。#7119は、医師、看護師、救急救命士から電話で助言を受け、必要に応じて119へつなぐ人間の相談経路である[19]。

AIの影響を、固定された条件にしない

AI、データ、アルゴリズムは、本人や現場専門職が直接書き換えられないことが多い。

生活に影響するのに、変えられない。説明されない。拒否できない。どの結果で停止するかも分からない。そうなれば、AIはケアの介入対象ではなく、固定された環境条件になってしまう。

必要なのは、モデルを一度作って終えることではない。何のために使うか、どの情報を扱うか、どこで人が確認するか、本人がどう訂正・拒否できるか、どの結果で停止するかを、導入前後に見直せることである。

本人がコードを書けなくても、優先順位、失敗例、訂正、異議、評価指標、停止条件へ影響できれば、AIの作用を部分的に変更可能な設計対象へ戻せる。

PreA研究は、患者、ケアパートナー、地域保健従事者、医療職を含む共創が、利用目的、言葉、失敗例、専門職への引き継ぎを調整する一つの経路になり得ることを示す。ただし、一つの共創研究だけで、参加型設計が危害を防ぐことや、通常のcodesign、human-centered AI、quality improvementより優れることは示せない。

AIの影響を、誰がどこまで理解すべきか

全員がモデル学習、統計、アルゴリズム監査を理解する必要はない。

本人と家族に必要なのは、理論名を説明できることではなく、次を知り、使えることである。

- いま応答しているのはAIか人か。
- 何が送信・保存されるか。
- AIが誤る可能性と、専門職へ確認すべき境界はどこか。
- 出力を訂正、拒否、停止できるか。
- AIを使わず、電話、紙、対面へ移れるか。

専門職はさらに、適用範囲、確認・上書き、説明、緊急離脱、生活への影響を理解する必要がある。

管理者・行政は、調達、責任、監査、データ統治、インシデント、格差、費用、停止・撤退を扱う。開発者・研究者は、データ来歴、モデル版、設定、性能分布、変更管理、比較設計を担う。

問題は、AIの影響を受ける本人に、変更を求める方法、拒否、停止、異議申立て、代替が伝わらず、実際に使えないことである。それは技術理解の不足だけでなく、知識翻訳と権利設計の失敗として評価すべきである。

現時点の利用境界

AIへ担わせやすいこと	人間の確認と組み合わせること	AIへ単独で担わせないこと
本人の言葉を短く整理する	診療前レポート、患者メッセージ下書き	個別診断の確定
質問候補を作る	診療録の平易化と誤り訂正	治療・薬・運動負荷の決定
経過や困る場面を時系列化する	選択肢と本人の希望の比較	緊急度の最終判断
専門職へ渡すメモのたたき台	一般情報と個別状況の照合	自傷他害・虐待・急変への危機対応
紙・電話・対面へ移る案内	データ送信・保存の説明	説明・承認・救済の責任

これは永久の線引きではない。新しい前向き研究、規制、監査、実装条件によって更新する必要がある。

まだ言えないこと

- AI相談が専門職を代替できる。
- 長く共感的な回答は正確で安全である。
- 免責文があれば緊急度判定は安全である。
- 特定モデルの精度を、別の版、prompt、地域、自由文相談へ一般化できる。
- Care Compassの緊急語検知は危機を漏れなく検出する。
- store:falseは匿名化、機密保持、Zero Data Retentionを保証する。
- 相談準備が健康、生活、作業参加を改善する。
- 共創を行えば、偏りと危害が解消する。

結論

AI相談は、専門職の代わりになれるのか。

現時点の答えは、代わりになれるとは言えないである。

ただし、何も任せられないわけではない。AIは、本人が話したいことを整理し、質問を作り、専門職へ渡すメモを準備し、専門職が確認する文章の下書きを作れる可能性がある。

その価値を引き出すには、AIの性能だけでなく、誰が目的を決め、どのデータを使い、どこで人が確認し、本人がどう訂正・停止し、どの結果で変更・撤退するかを設計しなければならない。

AIを安全に使う設計とは、モデルだけを賢くすることではない。

AIの作用を、本人と支援者が確認し、訂正し、拒否し、必要なら停止できるようにすること。

参加型設計がその仕事にどこまで役立つかは、今後、実利用者を含む比較研究で確かめる必要がある。

参考文献

[1] Chen SF, et al. LLM-assisted systematic review of large language models in clinical medicine. Nature Medicine. 2026;32:1152-. doi:10.1038/s41591-026-04229-5

[2] Johnson JT, et al. Research on patient-facing chatbots based on large language models in the care of older people: a living systematic review. European Geriatric Medicine. 2026;17(3):1557-1567. doi:10.1007/s41999-026-01454-6

[3] Ayers JW, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Internal Medicine. 2023;183(6):589-596. doi:10.1001/jamainternmed.2023.1838

[4] Tao X, et al. An LLM chatbot to facilitate primary-to-specialist care transitions: a randomized controlled trial. Nature Medicine. 2026;32(3):934-942. doi:10.1038/s41591-025-04176-7

[5] Li S, et al. A community-codesigned LLM-powered chatbot for primary care: a randomized controlled trial. Nature Health. 2026;1(2):238-. doi:10.1038/s44360-025-00021-w

[6] Draelos RL, et al. Large language models provide unsafe answers to patient-posed medical questions. npj Digital Medicine. 2026. doi:10.1038/s41746-026-02428-5

[7] Reis A, et al. Disclaimers and Referral Patterns for Medical Advice Across Urgency Levels: Large Language Model Evaluation Study. Journal of Medical Internet Research. 2026. doi:10.2196/84668

[8] Gao Y, et al. Accuracy of the large language model ChatGPT in adult emergency department triage: a systematic review and meta-analysis. BMC Emergency Medicine. 2026. doi:10.1186/s12873-026-01521-y

[9] Harada Y, et al. Safety Audit of a Large Language Model for Lay Self-Triage Using Japanese Symptom Vignettes. Cureus. 2026. doi:10.7759/cureus.105878

[10] Ramaswamy A, et al. ChatGPT Health performance in a structured test of triage recommendations. Nature Medicine. 2026. doi:10.1038/s41591-026-04297-7

[11] Pasli S, et al. ChatGPT-supported patient triage with voice commands in the emergency department: A prospective multicenter study. American Journal of Emergency Medicine. 2025. doi:10.1016/j.ajem.2025.04.040

[12] Sansoni JE, Grootemaat P, Duncan C. Question Prompt Lists in health consultations: A review. Patient Education and Counseling. 2015;98(12):1454-1464. doi:10.1016/j.pec.2015.05.015

[13] Ey C, et al. Improving communication and patient information recall via a question prompt list: randomized clinical trial. BJS. 2023. doi:10.1093/bjs/znad303

[14] Wang X, et al. Question prompt list intervention for patients with advanced cancer: a systematic review and meta-analysis. Supportive Care in Cancer. 2024. doi:10.1007/s00520-024-08432-3

[15] Vo MT, et al. Prompting Patients with Poorly Controlled Diabetes to Identify Visit Priorities Before Primary Care Visits: a Pragmatic Cluster Randomized Trial. Journal of General Internal Medicine. 2019;34(6):831-838. doi:10.1007/s11606-018-4756-4

[16] 個人情報保護委員会. 生成AIサービスの利用に関する注意喚起等. 2023年6月2日. 公式ページ: https://www.ppc.go.jp/news/careful_information/230602_AI_utilize_alert/

[17] OpenAI. Data controls in the OpenAI platform. 2026年8月23日確認. Official documentation: https://developers.openai.com/api/docs/guides/your-data/

[18] 厚生労働省. 「オンライン診療の適切な実施に関する指針」に関するQ&A. Q19-Q22. 公式ページ: https://www.mhlw.go.jp/web/t_doc?dataId=00tc8203&dataType=1

[19] 総務省消防庁. 救急安心センター事業（#7119）ってナニ？ 2026年8月23日確認. 公式ページ: https://www.fdma.go.jp/mission/enrichment/appropriate/appropriate007.html

[20] World Health Organization. Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models. WHO; 2024. ISBN 978-92-4-008475-9. Official publication: https://www.who.int/publications/i/item/9789240084759

利益相反・限界・更新

本稿はUnknown Labによる独立した対象限定レビューであり、査読済み系統的レビューではない。著者はCare CompassとETIC 2.0の開発者であり、自己評価上の利益相反がある。Care Compassの効果と安全性は第三者検証されていない。

PreA論文には、計算資源と技術支援を提供したTencent / WeChat AIの社員である共同著者が含まれる。P&P Care論文は競合利益なしと報告する。各研究の利益相反の有無だけで結果を採否せず、研究設計、対象、評価指標、再現性と合わせて判断した。

LLM、製品、prompt、データ管理、法令・指針は変化する。本稿は2026年8月23日までに確認した資料と実装に基づく。