論文とETIC 13

「専用AI」の上乗せは、何と比べて確かめるのか

製品の比較と、仕組みの効果を分ける



Daiki Morimoto / Unknown Lab
ORCID: 0009-0009-6444-750X
Version 1.0 / 公開日 2026-10-14
DOI: 10.5281/zenodo.23353130
未査読 / 専門職としての著者確認済み
独立専門レビュー未実施 / AI支援執筆
ライセンス: CC BY 4.0

30秒で読む

医療専用のAIが、汎用AIよりよい回答を返すとは限らない。Vishwanathらは、医学の試験問題、医療会話の評価課題、医師が実際に入力した質問で両者を比較した。医師の質問に対する平均評価は、調べた汎用AI群の方が高かった。ただし、これは特定の時期と条件での製品比較であり、「医療向けに特化すること自体が逆効果」という実験ではない。[1]

もう一つの研究では、AI回答の採点結果が、詳しく書かせるか短く書かせるかで大きく変わった。出力の長さを統計的に調整すると変動は小さくなったが、調整後の点数が正しさや安全性を直接示すわけではない。[2]

二つの研究から本稿が導くのは、比較の目的を先に分ける必要性である。「現場でどの製品が役立つか」と「同じ基礎モデルに加えた仕組みに効果があるか」では、そろえる条件が異なる。専門的な枠組みを加えたAIを評価する場合も、弱い汎用AIを相手に勝つだけでは、その枠組みの上乗せは確かめられない。

1. 詳しい回答が勝ったとき、何が勝ったのか

同じ相談内容を二つのAIに入力する。一方は短く要点を返し、もう一方は背景、関係者、確認事項を整然と並べる。後者を見て「こちらは状況をよく理解している」と感じることはある。だが、判断に必要な情報を増やしたのか、同じ情報を丁寧に見せたのか、単に読む量を増やしたのかは、最初の印象だけでは分からない。

これは架空の比較場面であり、研究の参加者や結果を再現したものではない。重要なのは、短い回答を優れたものと決めることでも、長い回答を疑うことでもない。注意事項を省略しないために長くなる回答もあれば、肝心な判断を曖昧にしたまま長くなる回答もある。文章の長さと内容の質は、別々に動くとは限らない。

例えば、ケアの予定を調整する場面で、汎用AIには「案を出して」とだけ頼み、専用AIには本人の希望、関係者への確認、代替案、未解決事項まで

整理するよう指示したとする。後者が高く評価されても、固有の理論が優れていたのか、単に指示が具体的だったのかを分けられない。比較相手にも、目的に合った十分な指示を与える必要がある。

逆に、実際の利用者がその指示を毎回作れないなら、最初から適切な出力を返す製品には実用上の価値がある。「指示の差だから価値がない」と切り捨てるのも誤りである。同じ観察から異なる結論が出るのは、評価しようとしている対象が違うからだ。製品の使いやすさと、追加した構成要素の効果を、一つの順位表で済ませないことが出発点になる。

比較したいこと	比較の基本	結果から言える範囲
実際に使う製品の有用性	製品ごとの通常の使い方を明示して比べる	その利用条件での製品全体の差
追加した仕組みの効果	同じ基礎モデルと資源で、その仕組みの有無を比べる	統制した条件での追加部分の差
長さに依存しない評価	元の出力と採点を保持し、長さ調整の感度を調べる	調整の仮定に依存する評価の変化

表1。比較の目的を分けるための本稿の整理。三つは代替関係ではない。特に、統計的な長さ調整を、仕組みの有無を割り付ける実験と同一視しない。

2. 医療専用AIと汎用AIをどう比べたのか

Vishwanathらの研究は2026年6月にNature Medicineへ公表された。先行する医学ベンチマーク評価に、医師の実利用から得た質問と、モデル名を隠した臨床家の採点を組み合わせた点が重要である。試験問題に答えられることと、医師が日々尋ねる問いに使えることを、同じものとして扱わなかった。[1]

公開ベンチマークは、MedQAとHealthBenchから各500項目を無作為抽出したものだった。MedQAは医学の選択式問題、HealthBenchは医療会話への回答を項目別の基準で採点する課題である。いずれも500人の患者を追

跡した研究ではない。HealthBenchの得点も、患者の治癒率や診断正答率を表す数値ではない。

医師の実質問は、NYU LangoneのHIPAAに準拠したGPT利用環境に由来する、匿名化された100件だった。これをreal-world clinical queries、以下RCQと呼ぶ。質問を集めた正確な期間や、回答を使った後の患者経過を観察した期間は、本研究からは確認できない。実利用に由来する質問であっても、患者への介入試験とは区別する必要がある。

公開ベンチマークの抽出には乱数シード62が示されている一方、RCQの母集団、質問の選定・除外過程は、それと同じ詳しさでは記述されていない。100件を無作為に選んだ代表標本と推定するべきではない。報告チェックリストでは標本規模に実務上の制約があったことが示されており、まれな害の同等性を検証するための標本設計とは区別する。

RCQでは、GPT-5.2、Gemini 3.1 Pro Preview、Claude Opus 4.6の汎用モデルと、OpenEvidence、UpToDate Expert AI、Google AI Overviewを比較した。12人の米国の臨床家が、各回答を3人ずつ評価した。正確性、網羅性、有害性の回避、明瞭性の4項目を各1〜4点で採点し、別に有害性と幻覚の有無を判定した。評価者には回答の提供元を知らせなかった。

評価対象	規模と採点	期間・比較条件
MedQA	無作為抽出500問題、選択肢の正答率	汎用3モデルと医療専用2製品
HealthBench	無作為抽出500課題、基準に沿う得点	同じ5システム、3つのAI採点器による評価
RCQ	100質問×6システム、各回答3人	主に2026年2月の出力、12人の臨床家が盲検評価
RCQの解析対象	拒否32回答を除く568回答、1,704評価	全6システムで回答がそろう74質問でも確認

表2。原著の研究構成。初期の600回答に対する評価枠は1,800件であり、1,800人の独立した参加者ではない。Google AI OverviewはRCQのみ。医療専用製品の反復取得には2025年9月または11月の時点も含む。[1]

汎用モデルはAPI経由、専用製品はブラウザ経由で利用された。API側では検索を有効にするなど設定をそろえたが、専用製品の内部モデル、検索、指示、更新過程の全てを同一にしたわけではない。公開コードも、乱数シードは対応する場合に設定するという条件を含む。設定値が書かれていることと、異なる製品間で完全に同じ推論条件を実現できることは別である。[1,5]

回答の長さや書式は、意図的に統一しなかった。実際の製品の説明の仕方も、臨床的な使いやすさの一部として残すためである。この選択は、製品全体を比べる目的には理解できる。一方、検索や専門化の仕組みだけの効果を問うには、基礎モデルを一致させた追加実験が必要になる。評価条件の長所と限界は、どちらの問いを立てるかで変わる。

3. 点数差は明瞭だが、安全性の優劣までは示さない

MedQAではGeminiが97.4%、GPTが94.2%、Claudeが90.2%、OpenEvidenceが89.6%、UpToDateが88.4%だった。HealthBenchでは、汎用3モデルの得点が77.0〜88.0、専用2製品は61.3〜62.6だった。両者を並べる際には、前者が選択式問題の正答率であり、後者が会話評価基準に基づく得点であるという分母の違いを残す必要がある。[1]

システム	MedQA正答率 % [95% CI]	HealthBench得点 [95% CI]
Gemini 3.1 Pro Preview	97.4 [95.6, 98.5]	79.3 [76.6, 81.9]
GPT-5.2	94.2 [91.8, 95.9]	88.0 [85.9, 90.1]
Claude Opus 4.6	90.2 [87.3, 92.5]	77.0 [74.2, 79.9]
OpenEvidence	89.6 [86.6, 92.0]	62.6 [59.3, 65.9]
UpToDate Expert AI	88.4 [85.3, 90.9]	61.3 [58.0, 64.6]

表3。各500項目の公開ベンチマーク。CIは信頼区間。異なる列の得点を同じ尺度として引き算しない。MedQAで全ての汎用・専用の組合せに有意差があったという意味ではない。[1]

MedQAは正答抽出をAIと正規表現で照合し、不一致を人が確認した上で、対応のある正誤をMcNemar検定で比較した。HealthBenchは3つのAI採点器の多数決を使い、得点差にはWilcoxon符号付順位検定を用いた。複数の組合せを調べるための補正も行っている。ただし、比較対象と採点器に同じモデル群が含まれ、複数AIの多数決を完全に独立した評価と見なすことはできない。[1]

RCQの平均評価は、Gemini 3.62、GPT 3.54、Claude 3.52に対し、OpenEvidence 3.24、UpToDate 3.17、Google AI Overview 3.27だった。汎用群と比較群の間の9組合せでは、多重比較を考慮しても差が報告された。同じ群の中で差が有意でないことは、各製品が同等だという証明ではない。とくに、平均差と実務上重要な差は別に判断する必要がある。

システム	RCQ平均 [95% CI]	回答数 / 拒否数
Gemini 3.1 Pro Preview	3.62 [3.56, 3.68]	98 / 2
GPT-5.2	3.54 [3.47, 3.61]	97 / 3
Claude Opus 4.6	3.52 [3.44, 3.59]	99 / 1
OpenEvidence	3.24 [3.17, 3.32]	99 / 1
UpToDate Expert AI	3.17 [3.09, 3.25]	81 / 19
Google AI Overview	3.27 [3.18, 3.35]	94 / 6

表4。100質問に対する平均評価と回答拒否。点数は拒否を除いた回答の評価であり、全100質問を通じた有用性の総合点ではない。各平均の尺度は1〜4点。[1]

拒否の扱いは特に重要である。評価者が拒否を指摘した回答は、研究者が確認して解析から除いた。UpToDateの拒否は19件で、他のシステムより多い。残った81件で良い回答を出す能力と、100件の依頼全体で役に立つ能力は一致しない。一方、危険な依頼を適切に断ることまで失敗扱いにするべきでもない。

原著は、全製品の回答がそろった74質問でも分析し、除外の影響が小さいという結果を示した。報告された平均値への最大の影響は0.034点だった。ただし、これは比較できた質問で順位の頑健性を調べる分析であり、拒否が妥当だったか、拒否後に人が何をしたかを評価したものではない。利用時の判断には、回答の質と適切な拒否を別々に残す方が分かりやすい。

また、二値の有害性指標ではp=0.55、幻覚指標ではp=0.42であり、群間差は有意でなかった。この結果から、どの製品も同じ程度に安全だとは言えない。100質問の評価は、まれでも重大な失敗の差を除外する非劣性試験ではない。平均点が高いシステムにも重大な失敗は起こりうるし、臨床家が文章を安全そうと採点することと、実際の患者に害が生じないことも異なる。

4. 「盲検評価」だけで採点の問題は消えるか

モデル名を隠したことは、この研究の長所である。それでも、長さ、語り口、引用の出し方から製品の特徴を推測できる可能性は残る。盲検は名前への先入観を減らす操作であって、回答に含まれる全ての手掛かりを消す操作ではない。書式を残すという研究目的との両立を考えれば、この限界を隠さないことが重要になる。

評価者間の一致にも注意が必要である。原著の順序尺度に対するKrippendorffのαは0.10〜0.20だった。1点以内の一致は89〜95%、二値化した指標では別の一致係数が高くなったが、これらは同じ問いへの答えではない。細かな4段階の違いについて意見が分かれることと、おおむね許容できる回答かどうかに合意することは両立する。[1]

原著は評価者の違いを固定効果やランダム効果で扱う追加解析も行った。この点は、採点のばらつきを無視した単純なランキングより丁寧である。ただし、同じ質問から複数システムの回答が作られ、その回答を複数人が採点する構造がある。1,704評価を1,704件の独立した臨床事例と見なすことはできない。将来の比較では、質問と評価者の両方に由来する依存を解析計画へ明示したい。

低評価の回答には、任意の自由記述も付けられた。著者らは101件の注記を帰納的に整理し、複数付与を許して153の工ラーラベルを得た。多かったのは不完全な回答55、安全上の省略33、事実の誤り20で、文章構成の悪さ18、読者に合わない冗長さ16なども含まれた。これは患者や臨床家への質的インタビューではなく、低評価時の任意注記の事後分類である。[1]

したがって、153件の有害事象が発生したとは読めない。注記のない回答に同じ問題がなかったとも言えない。それでも、正しさだけでなく「必要なことが抜けている」「読む相手に合っていない」という問題が評価を下げたことは、次の比較設計に有用である。長い回答を一律に削れば、安全上の省略を増やす可能性もある。単に文章量を減らすことが改善になるとは限らない。

費用と運用についても、研究は限定的である。拡張データには2026年4月時点のAPI従量料金、専門職向け無料サービス、有料購読などの提供形態が整理されている。しかし、実際の検索や推論に伴う全費用、待ち時間、引用の質、導入や監督に必要な人員、継続利用の負担までは比較されていない。料金表だけから費用対効果を結論づけることはできない。

さらに、この質問群は一つの医療機関に由来する。米国の医師が入力した問いへの評価を、日本の在宅ケア、多職種の調整、本人や家族による利用へ、そのまま移すことはできない。診断や薬剤の知識を問う回答が良好でも、本人の希望を確認する時機や、権限のない行動を止める能力が確かめられたわけではない。利用者と課題が変われば、比較の設計も変える必要がある。

5. 著者の結論と、公表後の異論

著者らの結論は、評価した条件では、汎用の先端モデルが専門的な臨床AIツールより高い成績を示したというものである。その意義は、「医療専用」という名称だけで性能を推定せず、独立した比較が必要だと実例で示した点にある。公開ベンチマークの汚染やAI採点器の偏りにも言及し、RCQを中心的な根拠として位置づけている。[1]

2026年9月には、Beaulieu-JonesとNematiが、この比較の結論を制約するベンチマークの限界を論じる書簡を公表した。公開冒頭では、RCQの対象範囲、推定の精度、回答生成条件の記述と統制に疑問を示している。著者の応答は、公開ベンチマークの限界を認めつつ、研究で用いた配置条件におけるRCQの比較結果を維持している。[3,4]

本稿で確認できた書簡と応答は、出版社が公開している冒頭部分までである。有料本文を読了したものとして、個々の批判が全て解消された、あるいは研究が否定されたとは述べない。また、このやり取りは撤回や訂正と同じではない。主論文の全文、拡張データ、報告チェックリストと公開コードを根拠とし、公表後に方法上の異論があることを付記する。

Unknown Labの読みは、著者の製品比較の結果を尊重しつつ、専門化という構成要素の因果効果へ結論を広げない、というものになる。専用製品の基礎モデルが非公開なら、差の原因は基礎モデルの能力、検索、指示、更新時期、画面上の制約など複数ありうる。「専門化は役立たない」も「汎用モデルで医療は十分」も、この設計からは導けない。

利益相反については、原著に企業持分やコンサルティング関係などの申告がある。これを結果が誤っている根拠にはしない一方、独立評価と再現可能性を読むための情報として扱う。RCQの元の質問データは公開されていないため、本稿では個票の再解析を行っていない。公開コードの存在と、第三者が同じデータで全解析を再現できることは分けて記録した。

6. 長く書くだけで、AIの順位は動く

比較条件の問題を、別の角度から具体化するのがDuboisらのLength-Controlled AlpacaEvalである。2024年のCOLMで発表され、本稿では2025年3月の著者公開版v2を確認した。対象は医療ではなく、一般的な指示に対するAI回答の自動評価である。新しい評価者を一から学習させるのではなく、既存の評価結果へ解釈可能な調整を加えた点に方法上の特徴がある。[2]

AlpacaEvalは、固定された805の指示について、評価対象モデルと基準モデルの回答をAI採点器が比較する。ここでの勝率は、採点器が対象モデルを好む確率を平均したものだ。805人の人間が正誤を判定したわけではなく、答えの事実性や安全性だけを測る指標でもない。対象となったモデル群や人間の順位との比較は、2024年当時の状況に属する。

研究では、同じモデルに詳しく書くよう指示した場合と、必要な情報を残して簡潔に書くよう指示した場合を比べた。GPT-4の一つの版では、元の勝率が22.9%から64.3%まで動いた。これはモデルを入れ替えなくても、出力の指示だけで大きく順位が変わりうることを示す。ただし、指示によって内容自体も変わりうるので、情報が完全に同じで長さだけを変えた実験とは言えない。

長さ調整後の同じ範囲は41.9%から51.6%だった。変動は小さくなったが、消えたわけではない。著者は、詳しく書かせるだけで評価を動かす操作への頑健性が改善したと解釈する。この結果は、回答を短くするほど正しくなるという主張ではなく、採点が文章量に強く左右される問題への対策として読むべきである。

指標・検証	元の評価	長さ調整後
詳細度の指示による勝率の範囲	22.9〜64.3%	41.9〜51.6%
人間による順位とのSpearman相関	0.94	0.98
長さ操作への感度：正規化標準偏差	26%	10%
悪い出力を切る攻撃の例	3.7%	12.2%：正則化あり

表5。Duboisらの主要結果。相関は共通する38モデルで比較。感度は原著Table 1の26%を採用し、本文の25%という記載との差を残す。攻撃例の正則化なしの調整値は25.9%であり、調整自体にも悪用への弱さがある。[2]

この表で、0.94から0.98への上昇を「統計的に有意な改善」と言い換えることには慎重であるべきだ。著者のブートストラップ検定では、元のAlpacaEvalとの比較はp=0.07、MT-Benchとの比較はp=0.06だった。通常の5%水準では有意とは判定しない。相関が高くなったという観察と、差についての不確実性を同時に残す必要がある。

さらに、比較先のChatbot Arenaも完全な正解ではない。利用者の好みには、読みやすさや表面上の説得力が含まれうる。人間の順位へ近づいたからといって、医療的な正しさや重大失敗を見分ける能力まで高まったとは言えない。一般的な回答の好みを近似する指標と、ケアの安全性を判断する指標は、目的も誤りの重みも異なる。

7. 長さをそろえることにも、別の落とし穴がある

長さ調整の中核は、モデルの違い、出力長の差、指示の難しさを説明変数とするロジスティック回帰である。出力長の差は標準化した後に変換し、非常に大きな長さの差が過度に効かない形にしている。5分割交差検証とL2正則化を用い、質問難易度を先に推定した上で各モデルの係数を求める。既存の採点結果に対する統計モデルである。[2]

その後、長さの差に対応する項をゼロにして、同じ長さだった場合の勝率を推定する。実際の回答を同じ文字数へ切り詰め、評価者にもう一度読ませたわけではない。反実仮想的な解釈は、モデル化と識別の仮定に依存する。観察された出力の長さを調整しただけで、全ての交絡が消えたと考えることはできない。

ケアの説明で、危険な行動を避ける注意書きが増えた結果、文章が長くなることを考える。この場合、長さは単なる装飾ではなく、有用な仕組みが働いた結果の一部かもしれない。長さを消す調整が、取り除きたい見かけの差だけでなく、残したい価値の経路まで除く可能性がある。何を効果と数えるかを決める前に、一律に長さを補正してはいけない。

逆に、同じ注意事項を何度も繰り返すことで高得点になるなら、長さの影響を点検する価値はある。ここでは、「長さをそろえれば公平になる」と「長さは価値だから全て残す」という二択を避けたい。まず通常の出力を保存し、内容の充足、重大な省略、読む時間を分けて評価する。その上で長さ調整を感度分析として加える方が、差の意味を追いやすい。

この二つの論文は、同じ結果を別々に確かめた研究ではない。医療論文は製品の自然な出力を残して臨床家が評価し、AlpacaEval論文は一般課題のAI採点に対する長さ調整を検討した。後者の係数を前者へそのまま当てはめることはできない。組み合わせる意味は、臨床ランキングを補正し直すことではなく、次の比較で何を統制し、何を成果として残すかを明らかにする点にある。

観察・操作	著者の解釈	本稿が越えない境界
汎用AIのRCQ平均が高い	評価条件で汎用モデルが優位	専門化そのものの因果効果ではない
詳細度の指示で勝率が変わる	評価は長さの操作に弱い	内容を固定した純粋な長さ効果ではない
長さ調整で変動が小さくなる	自動評価の頑健性が改善	臨床的な正しさや安全性の証明ではない
有害性指標に有意差がない	この標本では差を検出しない	安全性の同等性や非劣性を示さない

表6。原著の観察、著者の解釈、Unknown Labが設ける推論の境界。異なる研究の評価尺度を一つの効果量へ合成していない。

8. Unknown Labの批判：弱い比較相手では上乗せを測れない

第一の論点は、比較相手の作り方である。独自の枠組みを詳しく実装したAIを、曖昧な一文の指示しか与えていない汎用AIと比べれば、差が出ても不思議ではない。本人の希望、選択肢、未確認事項、権限の限界を整理する程度なら、既存の一般的な構造化指示でも実現できる可能性がある。これを比較から外すと、固有の理論が必要だったかを問えなくなる。

比較相手に十分な指示を与えることは、独自の理論を不利にする操作ではない。独自の部分が何を加えるかを明確にするための最低条件である。ただし、相手だけ無制限に改善することも公平ではない。両条件に与える開発時間、試行回数、参照する事例、評価結果へのアクセスをそろえ、評価用事例を使って片方だけ調整しない必要がある。

第二の論点は、資源と出力を混同しないことである。同じ出力上限を与えることと、完成した回答を完全に同じ長さにすることは違う。前者は利用可能な資源をそろえる設計であり、後者は介入の結果として生じた差を加工する操作になりうる。文字数を一致させるために重要な注意を消したり、

短い側へ無意味な文を足したりすれば、もはや元のシステムを比べていない。

第三の論点は、良い平均点の裏にある失敗を残すことである。説明が丁寧になった一方で、確認が必要な場面で行動を進める、本人の希望を推測で埋める、権限のない約束をする、といった失敗が増えるなら、単純な総合点は改善を誤って示しうる。これらは本稿が将来の評価に提案する失敗例であって、今回の原著がそのまま測定した項目ではない。

第四の論点は、評価方法にも比較相手が必要だということである。AI採点器の得点だけで結論を出さず、提供元を伏せた人間の判定と照合する。ただし、AIと人間が一致したことを査読と呼ばない。研究方法、医療・ケアの専門性、倫理、本人や家族の経験、著者としての責任は、それぞれ異なる確認である。一つの高性能モデルで代替できたと扱うべきではない。

9. ETIC 2.0への含意

考え方と重なる点

ETIC 2.0が技術の価値を人の生活や関係から捉えようとするなら、製品名や知識問題の点数だけで評価を終えない姿勢は、その考え方と重なる。[6] 誰が回答を読むのか、何を判断するために使うのか、どの失敗を避けたいのかを先に明らかにする必要がある。医療論文が明瞭性や網羅性を分けた点にも、この読み方との接点がある。

ただし、これらの研究はETIC 2.0を直接検証していない。長さの影響や比較条件の重要性は、既存の評価研究、実験計画、測定論から説明できる。既存の説明だけで十分なら、新しい概念を追加する必要はない。重なる点があるという評価を、有効性の証明や独自性の根拠へ置き換えないことが重要である。

答えるべき問い

独自の視点を加えたAIは、よく設計された一般的な構造化指示よりも、何を改善するのか。単に長く、丁寧に、専門的に見える文章を返すことではなく、同じ情報と資源の下で、必要な確認や判断の質を変えるのかが問いになる。その差が本人、家族、専門職の誰にとっての利益なのかも、一つの総合点に隠さず示す必要がある。

もう一つは、比較の目的に応じた成果を選べるかである。実用品としての評価なら、指示を作る手間、回答を読む時間、修正の負担も価値に含まれる。仕組みの検証なら、そうした差が基礎モデルや資源の違いから生じていないかを確かめる。二つの問いを分けた後で結果を照合する方が、単一の順位を万能の答えにするより透明である。

具体的な見直し案

本稿は、比較を二つの系統に分けることを提案する。第一は、同じ基礎モデル、同じ情報、同じ利用可能なツール、同じ出力上限と開発予算で、十分に作り込んだ一般的な指示と、検討したい追加要素を比べる試験である。第二は、実際の製品を通常の形で使い、手間や時間も含めた有用性を比べる試験である。両者の結果を混ぜて、追加要素だけの効果と呼ばない。

設計項目	最初に決めること	残すべき結果
比較相手	十分な一般的構造化指示と同じ開発予算	最弱の対照だけとの比較にしない
入力と資源	同じ事例・基礎モデル・情報・ツール・上限	版、設定、実行回数、失敗も記録
出力の長さ	上限を合わせ、実際の長さは改変せず保存	無調整評価を主に、調整は感度分析
回答拒否	妥当な拒否と不必要な拒否を区別	全依頼の分母と回答時の質を併記
採点	名前を隠し、質問と評価者の依存を扱う	重大失敗、主体別結果、一致の不確実性
実用性	通常の製品利用は別の比較として設計	読む時間、修正、費用、人員の負担

表7。Unknown Labによる未検証の比較設計案。原著の推奨事項を逐語的に移したものではなく、実験の実施や安全性を保証するものでもない。

この案の検証では、評価事例と重大失敗の基準を事前に定め、結果を見てから都合のよい指標へ変更しないことが重要になる。長さ調整の有無で結論が変わる場合は、どちらかを正解として選ぶのではなく、その違いを説明する。調整によって消えた差が、無用な冗長さだったのか、必要な内容の増加だったのかを、別の内容評価で調べる必要がある。

また、実験段階の提案を、そのまま現場で使う推奨に変えてはいけない。初期検証は、評価可能な模擬事例や公開可能な材料から始める選択肢がある。実際の本人や家族の情報を使う研究には、別途の倫理、同意、情報管理、責任分担の検討が必要である。今回の記事を完成させたことは、そのような実証や承認を済ませたことを意味しない。

10. 何が分かれば、上乗せという主張を弱めるか

十分に作り込んだ一般的な構造化指示との比較で、重大失敗や必要な判断に実用上意味のある差がなければ、独自の仕組みによる上乗せという主張を弱める。単に有意差がないという理由だけで同等と断定せず、推定の幅がどの程度の差をなお許すかを調べる。逆に、よい差があっても、それが長さや検索資源の差だけで説明できるなら、固有の仕組みの効果は保留になる。

平均点が上がっても、本人の選択を狭めるような重大失敗が増えるなら、総合的な改善という主張は維持できない。また、特定の採点AIだけが高く評価し、別の評価者や人間の盲検評価では差が消える場合には、評価器への適合を疑う。これは、結果を否定するための条件ではなく、どこまでなら主張してよいかを前もって決めるための条件である。

現時点で、二つの論文から特定のケアAIの優劣を決めることはできない。しかし、次の比較で混ぜてはいけないものは明確になる。製品としての有用性、追加した仕組みの効果、採点器が好む表現の三つである。この区別を実装し、なお差が残るかを調べることが、説明のよさを研究上の根拠へ近づける具体的な一歩となる。

11. 次に読む論文

次の候補は、Panickssery、Bowman、Fengによる「LLM Evaluators Recognize and Favor Their Own Generations」である。[7] 本稿では会議の公開抄録と書誌を確認した段階であり、数値的な根拠として採用していない。AI採点器が自分に由来する出力を認識し、好む可能性を扱う研究として、次の検討対象に位置づける。

今回残った問いは、長さを点検し、提供元の名前を隠しても、採点器と回答モデルの関係が評価を左右するか、という点である。読む際には、自己選好が成立する課題と条件、モデル間での違い、人間の評価との対応を確

かめたい。新たな採点モデルを増やすだけで独立性が得られると仮定せず、どの依存を減らせるかを検討する。

本稿は公開研究に基づく未査読の批判的論評であり、個別の診療判断や製品選定の推奨ではない。AI支援による調査・執筆を含み、独立した専門レビューは未実施である。著者はDaiki Morimoto、Unknown Lab。ORCID: 0009-0009-6444-750X。

参考文献

[1] Vishwanath K, Alyakin A, Ghosh M, et al. General-purpose large language models outperform specialized clinical AI tools on medical benchmarks. Nature Medicine. 2026;32:2405–2409. doi:10.1038/s41591-026-04431-5. 本文、Methods、Extended Data Tables 1–2、Reporting Summaryを確認。原著はCC BY-NC-ND 4.0。本稿の表は事実の独自整理であり、原著図版の転載ではない。

[2] Dubois Y, Galambosi B, Liang P, Hashimoto TB. Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators. COLM 2024. 著者公開版 arXiv:2404.04475v2, 2025-03-10. doi:10.48550/arXiv.2404.04475. 数値はv2のFigure 3、Table 1、ResultsとAppendixを確認。DOIはarXiv版の識別子。

[3] Beaulieu-Jones B, Nemati S. Limited benchmarks constrain the conclusions of a general-purpose versus clinical AI comparison. Nature Medicine. 2026-09-03. doi:10.1038/s41591-026-04638-6. 出版社公開冒頭のみ確認、有料本文未確認。

[4] Vishwanath K, Aphinyanaphongs Y, Oermann EK. Reply to: Limited benchmarks constrain the conclusions of a general-purpose versus clinical AI comparison. Nature Medicine. 2026-09-03. doi:10.1038/s41591-026-04637-7. 出版社公開冒頭のみ確認、有料本文未確認。

[5] NYU OLab. clinical-llm-benchmarks. 著者公開コードとREADME. 2026-10-14確認。コードの実行、非公開RCQデータの取得・再解析は行っていない。

[6] Morimoto D. ETIC 2.0. Unknown Lab. doi:10.5281/zenodo.21848456. 既公開の理論的背景として参照し、今回の原著による直接検証としない。

[7] Panickssery A, Bowman SR, Feng S. LLM Evaluators Recognize and Favor Their Own Generations. Advances in Neural Information Processing Systems. 2024;37:68772–68802. 会議公開抄録. arXiv版DOI:10.48550/arXiv.2404.13076. 次読候補、本文の定量的根拠には未採用。