資　料

臨床検査学教育 Vol.18, No.1 p.1 ～ 6, 2026.

マルチモーダル対応大規模言語モデルの臨床検査技師国家試験における正答率評価 ―テキストおよび画像問題の比較―

宿　利　　淳 *

要　旨　近年、大規模言語モデル（LLM）は医療分野においても急速に応用が進んでおり、特に画像入力に対応したマルチモーダル機能を備えた次世代モデルの性能向上が期待されている。これまでの報告では ChatGPT 4.0 の正答率は 70.8%、画像問題は 42.6% に留まっていた。本研究では画像問題を含む直近 5 回分の臨床検査技師国家試験を対象に，ChatGPT-4o、ChatGPT-o3、Google Gemini、Claude Sonnet 4 の 4 モデルをプロンプトなしで評価した。全体の正答率は 87.4 ～ 93.3%（平均 89.4%）、画像問題正答率は 58.7 ～ 79.3%（平均 68.6%）を示し、従来比でそれぞれ約 +19 パーセントポイント、+26 パーセントポイントの大幅な改善を認めた。一方臨床生理学および臨床免疫学領域では他分野に比べて得点がやや低く、画像認識の精度向上や専門的知識の追加学習、Chain-of-Thought の併用が精度向上に寄与すると考えられる。マルチモーダル AI / LLM が臨床検査の教育・実務支援に実用可能なレベルであるかを今後は検証する必要がある。

キーワード　大規模言語モデル、ChatGPT、Gemini、Claude、生成 AI、臨床検査技師国家試験

I. 緒　　言

近年の人工知能（Artificial Intelligence: AI）の進歩は著しく、とりわけ大規模言語モデル（Large Language Models: LLM）の登場は医学界にも大きなインパクトを与えている。2022 年に OpenAI 社が ChatGPT を公開して以降、GPT 4、Gemini、Claude など、マルチモーダル機能や推論性能を強化した次世代 LLM が相次いでリリースされ、医学教育や臨床応用への研究が加速している。

LLM の医学的知識を客観的に評価する手法として、各国の医療系国家試験問題を解答させ、その正答率を指標に能力を定量化するアプローチが広がっている。日本の医師国家試験においては GPT 3.5 で合格基準に届かなかったものの、GPT 4 以降のモデルでは合格点に達したことが報告されている[1)～3)]。臨床検査技師国家試験においても類似の研究が行われており、土井らは GPT 3.5 で正答率の平均が 51.4%、GPT 4 で 79.8% と大幅な向上を示した[4)]。さらに大重らは複数モデルを比較し、ChatGPT-3.5 で 49.5%、ChatGPT 4.0 で 72.0%、Meta 社 Llama2 で 35.7%、Google Bard で 46.8% と報告している[5)]。同研究では ChatGPT 4.0 による画像問題の正答率も評価され、平均 42.63% と報告している。

大重らの報告以降、様々なマルチモーダル対応

* 東京工科大学医療保健学部臨床検査学科　§ shukurias@stf.teu.ac.jp

LLMが登場しておりChatGPT-4o、-o3、Google Gemini、Anthropic Claudeなど画像入力に対応した新世代LLMが次々と公開された。しかし、先行研究の多くはテキストベースの問題に偏っており、血液塗抹標本や組織病理像といった臨床検査技師の日常業務および教育で不可欠な視覚情報を伴う問題に対する性能評価はほとんど行われていない。そこで本研究では直近5年間の臨床検査技師国家試験を対象に、ChatGPT-4o, ChatGPT-o3, Gemini 2.5 Flash, Claude Sonnet 4の4モデルを用いて、画像・非画像問題別の正答率、分野別の正答率を比較する。これにより最新LLMの画像読解能力を含めた性能を検証し、今後のAI発展と臨床検査領域での教育・実務利用の可能性について考察する。

II. 対象と方法

厚生労働省のホームページにて公開されている第67回（2021年）から第71回（2025年）の5回分の臨床検査技師国家試験問題[6)～10)]を解析対象とした。試験科目は厚生労働省が定める「臨床検査技師国家試験出題基準と試験科目との対応表」に基づいて臨床検査総論、臨床検査医学総論、臨床生理学、臨床化学、病理組織細胞学、臨床血液学、臨床微生物学、臨床免疫学、公衆衛生学、医用工学概論の10分野に分類した。問題および選択肢は改変せず、各LLMのWebページ上にコピー&ペーストして解答を得た。なお追加プロンプトは一切入力していない。各LLMからの解答の生成は2025年5月25日から2025年6月13日にかけて実施した。正答率の可視化はGraphPad Prism（GraphPad Software, San Diego, CA, USA）、単語の頻度分析はMATLAB R2024a（MathWorks, Natick, MA, USA）を使用した。全体正答率および画像問題の正答率はモデル間で一元配置分散分析およびTukeyの多重比較検定を行った。$P < 0.05$を有意水準とした。単語の頻度分析は統計的検定を行わず、記述統計として提示した。

III. 結　　果

A. 問題数の概要と解答形式

表1に各回の文章問題、画像問題、複数選択問題の数を示した。各回で画像問題は30問前後、複数選択問題は35問前後出題されていた。各モデルの生成した解答はいずれも正答番号を明示し、正答の根拠を提示する形式で一致していた。加えてChatGPT-4o、Gemini 2.5 Flash、Claude Sonnet 4では選択肢ごとの簡単な解説を併記した。

B. モデルごとの正答率

図1にChatGPT-4o, ChatGPT-o3, Gemini 2.5 Flash, Claude Sonnet 4それぞれのモデルにおける、過去5回の臨床検査技師国家試験の平均正答率および全体、画像問題の正答率比較を示す。国家試験問題全体での正答率はいずれのモデルにおいても87%を超え、ChatGPT-o3に関しては93%と最も高かった。モデル間で正答率に有意差を認め、ChatGPT-o3がChatGPT-4oとClaude Sonnet 4よりも有意に正答率が高値であった（$p < 0.01$）。文章形式の問題や複数選択肢の問題においてはいずれも90%の正答率を示した。一方で画像問題に関しては最も高いChatGPT-o3で79.1%、最も低いClaude Sonnet 4では58.2%となった。統計解析の結果、Claude Sonnet 4はChatGPT-o3よりも有意に正答率が低く（$p < 0.05$）、画像読解がモデル性能のボトルネックで

表1　各回の問題形式数

回	総問題数	文章問題	画像問題	複数選択問題
67	200	174	26	43
68	200	163	37	32
69	200	168	32	46
70	200	170	30	30
71	200	167	33	24

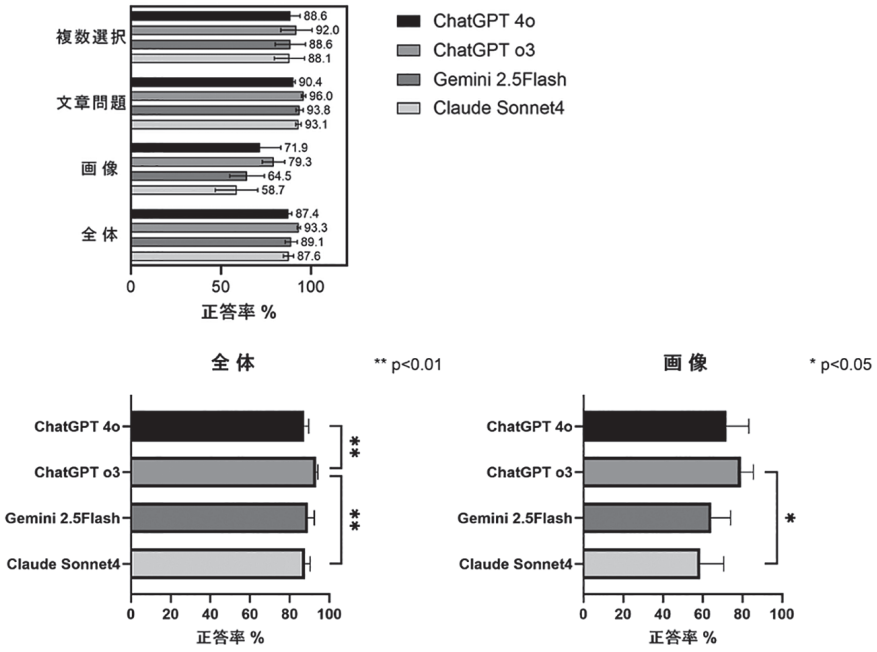

図 1　問題形式別の正答率（平均± SD，n=1250）（上段）
およびモデル間における全体および画像問題の正答率の比較（下段）

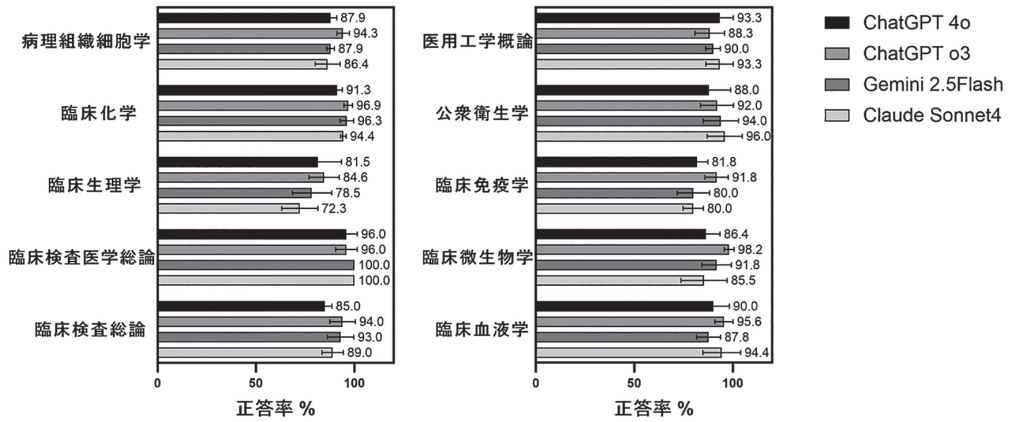

図 2　分野別正答率（平均± SD）

あることが明らかになった。なお、各年度ごとの正答率の推移については本文中で個別には示していないが、過去 5 回の平均値を用いた解析結果（**図 1**）と整合しており、年度間で大きなばらつきは認められなかった。

C．分野別の正答率

図 2 に分野別の正答率を示した。多くの分野で 90% 以上の正答率を示す一方で、4 種類のモデルの平均値として臨床生理学が 79.2 ± 4.6%、臨床免疫学が 83.4 ± 4.9% と相対的に低値を示した。そこで両分野で誤答した問題の性質を検討するために、問題文および選択肢のテキストマイニングにより単語を頻度分析した（**表 2**）。

IV. 考　　察

本研究は、画像問題を含む臨床検査技師国家試験を用いて最新 LLM を横断比較した初の報告である。先行研究では ChatGPT 4.0 の総合正答率は 70.8 ～ 79.8%、画像問題の正答率は 42.6% にとどまっていた 4,5)。これに対し本研究で評価した 4 つの最新 LLM では総合正答率 87 ～ 93%（平均 89.4%）、画像問題正答率 58 ～ 79%（平均 68.6%）を達成し、総合で +19 パーセントポイント、画像で +26 パーセントポイントの大幅な性能向上を示した。ChatGPT-o3 は Chain-of-Thought（思考の連鎖）と呼ばれる、LLM の推論能力を向上させ説明可能性を高めるプロンプトエンジニアリング手法を用いており、画像処理技術を複合させて推論している [11]。ChatGPT-o3 は推論チェーン内で画像をズーム / クロップしながら再解析する機構を備え、不鮮明な部位の再読み込みにより視覚情報を精緻化でき、この高度な推論が正答率向上に奏功した可能性が高い。この一方で Claude Sonnet 4 での低迷は、解像度制限により画像が縮小される点や限定的な空間認識能力が原因と考えられる。公開されている開発者向けドキュメント [12] によれば、Sonnet 4 は長辺 1,568 ピクセルを超える画像を自動縮小して処理する。また軽量高速モデルであるため画像認識性能よりも応答速度を優先していることも一因であると推察される。実際に第 71 回午前の問題 17 では、筆者は 2,174 × 1,354 ピクセルの PNG 形式の画像を各モデルにアップロードしたが、正答した ChatGPT-4o、o3 では 2,048 × 1,275 ピクセルの約 11% 圧縮された PNG 画像を基に解答を生成した。一方で誤答した Gemini および Claude が処理した画像はそれぞれ 1,024 × 637 ピクセルの JPEG 形式、1,394

表 2　臨床生理学（A）および臨床免疫学（B）の誤答した問題での単語出現回数

（A）臨床生理学での出現回数

画像問題で使用される単語が多くを占めており、その後には波形の判断を要求する単語が続いている。

ChatGPT-4o（n=24）		ChatGPT-o3（n=20）		Gemini（n=28）		Claude（n=36）	
単語	回数	単語	回数	単語	回数	単語	回数
を別に示す	17	を別に示す	18	を別に示す	25	を別に示す	29
別冊 No	16	別冊 No	17	別冊 No	24	別冊 No	27
2 つ選べ	4	右脚ブロック	3	Stage	5	度房室ブロック	5
標準 12 誘導心電図	3	標準 12 誘導心電図	3	三相波	3	右脚ブロック	4
脳波	3	脳波	3	標準 12 誘導心電図	3	左脚ブロック	3

（B）臨床免疫学での出現回数

画像問題で使用される単語が半数を占めており、その後に輸血・移植検査領域の単語が続いている。

ChatGPT-4o（n=20）		ChatGPT-o3（n=9）		Gemini（n=22）		Claude（n=22）	
単語	回数	単語	回数	単語	回数	単語	回数
低値	8	別冊 No	4	別冊 No	10	別冊 No	11
基準範囲内	7	dL 以下	2	低値	8	を別に示す	8
別冊 No	6	Donath	2	基準範囲内	7	低値	8
抗 C	5	を別に示す	2	を別に示す	6	基準範囲内	7
HLA クラス	4	抗 C	2	抗 C	6	抗 C	6
を別に示す	4	抗 E	2	抗 E	4	抗 E	5
抗 Dia	4	提供者が O 型	2	16	3	2 つ選べ	4

× 868 ピクセルの webp 形式に圧縮されていた。Gemini では約 80%、Claude では約 40% の画素情報が圧縮されており、細部の情報が失われたことにより正答率が低下したと考えられる。

分野別では、臨床生理学と臨床免疫学で相対的にスコアが低下した。**表 2** より、臨床生理学ではいずれの LLM でも「を別に示す」、「別冊 No」という画像問題に関する単語が頻出しており、前述の画像認識性能がスコア低下に直結したと考えられる。一方臨床免疫学では「別冊 No」、「を別に示す」の他に「抗 C」や「抗 E」など輸血・移植検査領域の単語が頻出したため、画像認識性能以外に輸血・移植領域の学習が不足している可能性が示唆された。なお「基準範囲内」および「低値」は第 71 回午後問題 82 の試験管内補体寒冷活性化現象の検査結果の組み合わせを問う問題の中で多用されている。ChatGPT-4o、Gemini、Claude でこれらの単語が多く出現したのはこの 1 問を誤答したためである。他国の報告[13)14)]においても免疫学領域の正答率が低いことが報告されており、Yikai らは疾患の多様性と複雑さや、訓練データにおける情報の偏りや不足が原因であると指摘している。臨床応用や教育への利用を考慮すると、今後は専門用語の追加学習に加え、Chain-of-Thought といった高度な推論が不可欠である。

今後 LLM/ マルチモーダル AI は、臨床検査のワークフロー全域に段階的に浸透すると考えられる。検査結果を臨床検査情報システム LIS から直接取り込むことが可能となれば、生化学検査や血液検査、微生物検査、生理検査といった複数の検査領域を横断的に統合し、得られた全検査値を解釈したレポートを自動生成できる。従来のシステムでは各領域ごとに個別のコメントしか出力できなかったのに対し、LLM は異常値の組み合わせや相関から総合的な評価を提示できるため、検査レポートの質と効率が飛躍的に向上すると期待される。またガイドライン改訂や機器更新に伴う標準作業手順書（SOP）の改定にも LLM が利用できる。改定内容を LLM に読み込ませることで差分要点を抽出して SOP の自動更新も可能となる。

さらに AI は学習教材の自動生成にも応用でき、新人技師が苦手分野をピンポイントで復習できる個別最適化学習に利用できる。文章問題での高精度を活かし、模擬試験生成や国家試験・認定試験対策チャットボットとしての活用が期待できる。今後の AI はヒューマンエラーを最小化しつつ、高度な判断支援と業務効率化を両立する次世代プラットフォームへと進化していくと展望される。

本研究では画像は国家試験問題の PDF ファイルをキャプチャした静止画のみを検証対象としたため、DICOM 画像や動画については性能を評価していない。また、バージョン更新に伴う再現性問題、多言語対応への一般化可能性は今後の課題である。本研究は日本語の国家試験を対象としたため、他言語や他文化圏、さらには異なる形式の医療資格試験に対する一般化の可能性については未検討である。国際的には米国医師国家試験（USMLE）を対象とした複数の英語文献が存在し、ChatGPT をはじめとする LLM が高い正答率を示すことが報告されている[15)～17)]。これらとの比較や位置づけを行うことで、本研究の成果を国際的な文脈の中で評価することが可能となる。また言語依存性や文化的背景、学習データの偏りが LLM の性能に与える影響については十分解明されておらず、今後の重要な検討課題である。更に、単なる正答率の報告にとどまらず、解答仮定の妥当性や臨床教育への応用可能性、教育的効果の検証など、一歩踏み込んだ評価が必要である。

V. 結　　論

最新 4 種類のマルチモーダル LLM は臨床検査技師国家試験で 85% 以上の正答率を示し、ChatGPT-o3 は 93% と最も優れた性能を示した。画像問題ではモデル間で最大 21 パーセントポイントの差が残ったものの、ChatGPT-o3 は 79% に達し従来モデルを大きく上回った。LLM は臨床検査領域における教育・実務支援へ応用可能なレベルに近づきつつあるが、画像読解を中心とした課題克服が不可欠である。

投稿論文における COI 状態の開示

投稿論文に関連し、発表者らに開示すべき COI

関係にある企業などはありません。

文　献

1) Kasai J, Kasai Y, Sakaguchi K, Yamada Y, Radev D. Evaluating GPT-4 and ChatGPT on Japanese Medical Licensing Examinations. arXiv 2023. [Internet] Available from: https://arxiv.org/abs/2303.18027 [cited 2025 Jun 25]
2) Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: Comparison study. JMIR Med Educ 2023; 9: e48002.
3) Miyazaki Y, Hata M, Omori H, Hirashima A, Nakagawa Y, Eto M, et al. Performance of ChatGPT-4o on the Japanese medical licensing examination: Evalution of accuracy in text-only and image-based questions. JMIR Med Educ 2024; 10: e63129.
4) 土井洋輝 , 石田秀和 , 永沢大樹 , 坪井良樹 , 菊地良介 , 市野直浩 , その他 . ChatGPT による臨床検査技師国家試験正答率の検証 . 医学検査 2024; 73: 323-31.
5) 大重舞奈 , 二宮惇 , 赤座実穂 , 河原智樹 , 角勇樹 . 大規模言語モデルは臨床検査技師国家試験に合格することができるか . 臨床検査学教育 2024; 16: 99-105.
6) 第 67 回臨床検査技師国家試験問題および正答について , 厚生労働省 , 2021. https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp210416-07.html
7) 第 68 回臨床検査技師国家試験問題および正答について , 厚生労働省 , 2022. https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp220421-07.html
8) 第 69 回臨床検査技師国家試験問題および正答について , 厚生労働省 , 2023. https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp230524-07.html
9) 厚生労働省 . 第 70 回臨床検査技師国家試験問題および正答について , 厚生労働省 , 2024. https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp240424-07.html
10) 厚生労働省 . 第 71 回臨床検査技師国家試験問題および正答について , 厚生労働省 , 2025. https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp250428-07.html
11) OpenAI. Developer blog [Internet]. Available from: https://openai.com/ja-JP/index/introducing-o3-and-o4-mini [cited 2025 Jun 25].
12) Anthropic. Developer Guide [Internet]. Available from: https://docs.anthropic.com/en/docs/build-with-claude/vision [cited 2025 Jun 25].
13) Alfertshofer M, Knoedler S, Hoch CC, Cotofana S, Panayi AC, Kauke-Navarro M, et al. Analyzing Question Characteristics Influencing ChatGPT's Performance in 3000 USMLE®-Style Questions. Med Sci Educ 2024; 35: 257-67.
14) Chen Y, Huang X, Yang F, Lin H, Lin H, Zheng Z, et al. Performance of ChatGPT and Bard on the medical licensing examinations varies across different cultures: a comparison study. BMC Medical Education 2024; 24: 1372.
15) Kung TH, Cheatham M, Medenilla A, Sillos C, Leon LD, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2023; 2: e0000198.
16) Bicknell BT, Butler D, Whalen Sydney, Ricks J, Dixon CJ, Clark AB, et al. ChatGPT-4 Omni Performance in USMLE Disciplines and Clinical Skills: Comparative Analysis. JMIR Med Educ 2024; 10: e63430.
17) Nori H, King N, McKinney Scott Mayer, Carignan D, Horvitz E. Capabilities of GPT-4 on Medical Challenge Problems. arXiv 2023 ; 2303. 13375. [Internet]. Available from: https://arxiv.org/abs/2303.13375 [cited 2025 Aug 12]