原　著

臨床検査学教育 Vol.16, No.2 p.99 ～ 105, 2024.

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

大重舞奈*1　二宮　惇*1　赤座実穂*
河原智樹*　角　勇樹*§

要　旨　AIの大規模言語モデル（LLM）は、質問に対して人間のように自然な言語を生成することが可能である。本研究ではChatGPT3.5、ChatGPT4.0、Llama2、Bardの4種類のLLMモデルを用いて、画像問題を含めた臨床検査技師国家試験を解かせての正答率を初めて調べた。その結果、ChatGPT3.5で0.495、ChatGPT4.0で0.720、Llama2で0.357、Bardで0.468の正答率でChatGPT4.0のみが合格点に達した。AIは臨床検査技師国家試験に合格する知識を保有している。今後の臨床検査技師教育においても変革が求められ、将来AIを利用することを前提とし、AIの原理や活用法、ハルシネーションに代表されるAIの限界などの教育が必要と考える。

キーワード　LLM（Large Language Model）大規模言語モデル、ChatGPT、Llama2、Bard

緒言（目的）

最近のAIの進歩は医学界の変革をもたらし、医療人が行っていた多くの行為がAIに置き換わると思われる。これまでのAIの医療応用はパターン認識が主体であり、X線やCT診断などの放射線学[1]、病理学画像診断[2]、検査所見の解釈[3]が主体であった。しかし2022年に出現したOpenAIのChatGPTや、GoogleのBardをはじめ多くの大規模言語モデル（Large Language Models/LLM）は、大量のデータとディープラーニング（深層学習）技術により構築された言語モデルであり、人間のように質問への自然な回答を生成することができるため社会に大きな衝撃を与えた。これまでの研究ではLLMの医学的能力は、医師国家試験[4]と看護師国家試験[5]で画像無しの文書問題のみで評価されていた。臨床検査技師国家試験問題で検討したものは我々が検索した限りではなかった。そこで、画像問題を含めた臨床検査技師国家試験問題をLLMに解答させ正答率を検討した。

本研究ではAIの臨床検査技師国家試験回答能力を評価し、今後のAI発展と臨床検査技師教育の方向性の考察を行った。

I. 対象と方法

1. 使用したLLMの特徴

使用したLLMはChatGPT3.5、ChatGPT4.0、Llama2、Bardの4種類である。ChatGPT3.5[6]（Chat

* 東京医科歯科大学大学院医歯学総合研究科生命理工医療科学専攻生体検査科学講座生命情報応用学分野
§ sumi-alg@umin.ac.jp

1 この2人の著者は本研究に等しく貢献した

Generative Pre-trained Transformer）は、OpenAIが2022年11月に公開したモデルである。無料で公開されているChatGPT3.5ではテキスト入力のみ使用可能だが、有料版のChatGPT4.0では画像入力にも対応している。Llama（Large Language Model Meta AI）は、2023年2月にMeta AIから発表された。その後、2023年7月にLlamaの改良版としてLlama2[7]が公開された。Llama2は、少ないパラメーター数で高精度の回答が可能である。パラメーター数は多くなるほど高精度の回答が期待できるが、計算コストがかかるというデメリットがある。70億（7B）、130億（13B）、700億（70B）のパラメーターで学習した3つのモデルが選択可能であり、今回は最も高性能な700億パラメーターモデルを使用した。また、Llama2も画像入力に対応している。Bard[8]は、2023年2月にGoogleから公開されたモデルで、2024年2月8日に名称がGeminiに変更となっている。Google検索と連動して回答を作成しているためリアルタイムの情報を取得可能で画像入力にも対応している。

2．入力データ

本研究では、2020年（第66回）[9]、2021年（第67回）[10]、2022年（第68回）[11]、2023年（第69回）[12]の4年分の臨床検査技師国家試験問題を入力データとした。各LLMのWebホームページ上に試験問題をペーストし解答を得た。特段の設定変更は行なっておらずデフォルトのままで、プロンプトエンジニアリングは行っていない。試験問題は1年分で、午前100問と午後100問の計200問である。画像入力が可能なChatGPT4.0、Llama2、Bardについては画像問題の入力を行ったが、画像入力に対応していないChatGPT3.5ではテキスト問題のみの入力を行った。なお、複数回答が必要な問題では、問題文の通り「2つ選択して下さい。」と指示に明記した。

3．分析方法

4種類のLLMから得た回答に対し、厚生労働省から公表されている「臨床検査技師国家試験問題および正答について」を参照し採点を行った。採点結果で60%以上の正答率を得られた場合、合格と判定した。また、LLMの得手不得手を調べるために、単一回答と複数回答や試験科目ごとでも正答率の比較を行った。試験科目は全10科目に分けた（**表1**）。

II．結　　果

1．入力データの統計

4年分（第66回～第69回）の臨床検査技師国家試験問題を各LLMに示し回答を得た。実際に入力し回答を得ることが出来た問題数とその問題の形式は以下の通りである（**表2**）。ChatGPT3.5は画像入力に対応していないため、回答はテキスト問題のみの結果となった。Llama2、BARDは画像入力に対応しているが、「処理不能」という表記になり、回答を得ることが出来ない問題があった。ChatGPT4.0では全ての問題に対する回答出力が可能であった。

2．正 解 率

各LLMの正答率を**表3**に示した。最も正答率が高いものはChatGPT4.0で平均71%、最も低いものはLlama2で平均28%であった。臨床検査技師国家試験の合格ラインは60%であるため、今回利用したLLMではChatGPT4.0のみ合格となった（**図1**）。

表1　試験科目と問題数

分野	問題番号	問題数
臨床検査総論	1～10	10
臨床検査医学総論	11～15	5
臨床生理学	16～28	13
臨床化学	29～44	16
病理組織細胞学	45～58	14
臨床血液学	59～67	9
臨床微生物学	68～78	11
臨床免疫学	79～89	11
公衆衛生学	90～94	5
医用工学概論	95～100	6

臨床検査技師国家試験問題を厚生労働省から公開されている出題基準に基づいて分野ごとに分類した。出題傾向は臨床検査医学総論と公衆衛生学が最小で各5問、臨床化学が最大で16問である。

表 2　回答を得られた問題数

LLM	年	全形式問題	画像問題	計算問題	複数回答問題
ChatGPT3.5	第 66 回午前	83	0	1	24
	第 66 回午後	85	0	4	18
	第 67 回午前	87	0	2	19
	第 67 回午後	87	0	7	23
	第 68 回午前	83	0	4	17
	第 68 回午後	79	0	3	15
	第 69 回午前	82	0	1	25
	第 69 回午後	81	0	3	18
ChatGPT4.0	第 66 回午前	100	15	1	24
	第 66 回午後	100	15	4	18
	第 67 回午前	100	14	2	19
	第 67 回午後	100	11	7	23
	第 68 回午前	100	17	4	17
	第 68 回午後	100	21	4	15
	第 69 回午前	100	17	3	25
	第 69 回午後	100	19	3	20
Llama2	第 66 回午前	98	15	1	24
	第 66 回午後	94	12	4	18
	第 67 回午前	99	13	2	19
	第 67 回午後	92	10	7	23
	第 68 回午前	92	9	4	17
	第 68 回午後	86	8	4	15
	第 69 回午前	82	0	1	25
	第 69 回午後	82	0	3	18
Bard	第 66 回午前	98	15	1	24
	第 66 回午後	98	13	4	18
	第 67 回午前	98	13	1	19
	第 67 回午後	97	10	5	23
	第 68 回午前	98	15	4	17
	第 68 回午後	96	17	4	15
	第 69 回午前	97	14	3	25
	第 69 回午後	95	15	2	20

各 **LLM** に 4 年分の臨床検査技師国家試験問題を入力データとして与え、回答が出力可能であった問題数を示した。問題形式（画像問題、計算問題、複数回答が求められている）による回答出力可能問題数も記載した。**ChatGPT3.5** は、画像入力非対応であるため、回答問題数は **0** となっている。**ChatGPT4.0** のみが全ての問題に対する回答出力が可能であった。

表 3　ChatGPT3.5、ChatGPT4.0、Llama2、Bard の正答率

LLM	第 66 回正解率	第 67 回正解率	第 68 回正解率	第 69 回正解率	平均	SD
ChatGPT 3.5	53.57 %	58.05 %	48.77 %	50.31 %	52.67 %	3.56 %
ChatGPT 4.0	**68.50** %	**70.50** %	**71.50** %	**72.50** %	**70.75** %	1.48 %
Llama2	22.40 %	17.80 %	26.97 %	44.51 %	27.92 %	10.11 %
Bard	43.88 %	46.67 %	43.30 %	50.26 %	46.03 %	2.76 %

ChatGPT3.5、ChatGPT4.0、Llama2、Bard の正答率を示した。合格ラインである **60%** を超えている正答率は太字で示した。**ChatGPT4.0** のみ **4** 年分すべてで合格ラインに達している。

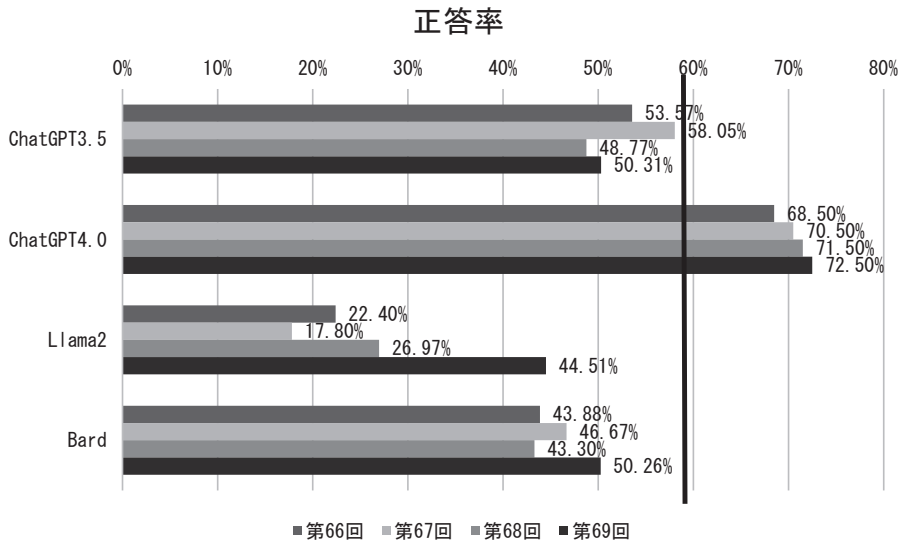

図 1　正答率のまとめと合格ライン

図 **1** は､表 **3** で示した各 **LLM** の正答率をグラフ化したもので､合格ラインである **60%** には太線を引いている。

ChatGPT3.5：第 **69** 回で、入力して正常に出力が返ってきた問題 **163** 問中正解した問題は **82** 問であった。第 **68** 回は **166** 問中 **79** 問正解であった。正解率は第 **69** 回で **50.3%**、第 **68** 回で **48.8%** とどちらも合格ラインである **60%** に達していない。**ChatGPT4.0**：第 **69** 回は、**200** 問中正解した問題は **145** 問であった。第 **68** 回は **200** 問中 **145** 問正解であった。正解率は第 **69** 回で **72.5%**、第 **68** 回で **71.5%** とどちらも合格ラインである **60%** を超えている。

Llama2：第 **69** 回で、**164** 問中正解した問題は **73** 問であった。第 **68** 回は **178** 問中 **48** 問正解であった。正解率は第 **69** 回で **44.5%**、第 **68** 回で **27.0%** とどちらも合格ラインである **60%** に達していない。

Bard：第 **69** 回は、**191** 問中正解した問題は **96** 問であった。第 **68** 回は **194** 問中 **80** 問正解であった。正解率は第 **69** 回で **50.3%**、第 **68** 回で **43.3%** とどちらも合格ラインである **60%** に達していない。

3．解答の特徴

LLM から得た解答の正答率を分野毎に分けて比較した結果を**表 4** に示した。ChatGPT4.0 は 10 分野中 9 分野で合格ラインである 60% を超えており、特に臨床検査医学総論に関しては、平均 87.5% と高い正答率となった。公衆衛生学はどのモデルでも正答率が低い傾向にあった。**表 5** に ChatGPT4.0 から得た回答の画像問題の正答率を示した。第 69 回午前のみ 82.35% と合格ラインである 60% を超えているが、他の回では 60% を下回る結果となった。

表 4　分野別の正答率

分野	chatGPT3.5	chatGPT4.0	Llama2	BARD
臨床検査総論	38.24 %	67.50 %	25.33 %	46.84 %
臨床検査医学総論	82.05 %	87.50 %	30.00 %	55.00 %
臨床生理学	49.25 %	64.42 %	17.86 %	42.55 %
臨床化学	48.82 %	77.34 %	33.07 %	44.53 %
病理組織細胞学	51.69 %	67.86 %	24.53 %	42.34 %
臨床血液学	42.86 %	65.28 %	20.00 %	40.28 %
臨床微生物学	59.72 %	76.14 %	30.49 %	54.55 %
臨床免疫学	64.00 %	68.18 %	26.83 %	34.09 %
公衆衛生学	36.84 %	57.50 %	25.00 %	46.15 %
医用工学概論	62.79 %	79.17 %	32.61 %	65.22 %

表 5　ChatGPT4.0 による画像問題の正答率

	問題数	正答数	正答率
第 66 回午前	15	2	13.33 %
第 66 回午後	15	3	20.00 %
第 67 回午前	14	5	35.71 %
第 67 回午後	11	2	18.18 %
第 68 回午前	17	8	47.06 %
第 68 回午後	21	11	52.38 %
第 69 回午前	17	14	**82.35** %
第 69 回午後	19	10	52.63 %
平均	129	55	42.63 %（ ± 21.68 %）

ChatGPT4.0 が回答した画像問題の正答率を示したものである。合格ラインである 60% を超えている正答率は太字で表した。ChatGPT4.0 は全ての画像問題において回答を出力している。最も高い正答率だったのは第 69 回午前で 82.35% だった。

III. 考　　察

今回は 4 年分の臨床検査技師国家試験問題を、ChatGPT3.5、ChatGPT4.0、Llama2、BARD の 4 種類の LLM に回答させた。現段階で合格ラインを超えることが出来たのは ChatGPT4.0 のみであった。得られた回答の中には、単一回答を求められている問題で複数回答をしたため不正解となった問題があった。Bard で、単一回答問題に複数回答を返した問題数は、第 66 回で 40 問、第 67 回で 32 問、第 68 回で 39 問、第 69 回で 25 問だった。この複数回答の中には正解の選択肢も含まれていたため、正確に選択肢を絞ることが出来ればより良い正答率が得られる可能性がある。また、画像問題に関しては、画像入力に対応している LLM であっても回答を出力する過程でエラーになり「処理不能」となる結果になるものがあった。ChatGPT4.0 では全ての画像問題に対して回答を得られたが、Llama2 では 74 問中 57 問（77.03%）、Bard では 13 問（17.57%）で処理不能となった。総

合点で合格ラインの 60% を超えている ChatGPT4.0 についても、画像問題においては 58.11% と 60% を下回る結果となり、LLM の画像処理は発展途上であると考えられる。今回使用した LLM は一般的な事項を学習したモデルであり、医療に特化したものではない。LLM は上書き学習を行って各利用者に特化したモデルにすることが可能である。多量の医療文献の上書き学習を行えば、より良い結果が得られることが期待される。LLM とは言語モデルのうち「計算量」「データ量」「モデルパラメータ数」を大規模化したもので、パラメータが 100 億を超えるモデルを LLM とすることが多い。モデルが一定以上大きくなると、急激に性能が向上する[13]。モデルパラメータ数は GPT3.5 5,350 億、GPT-4.0 は非公開であるが 1.7 兆、Llama-2 70 億、Bard 3,400 億であり、GPT-4.0 の高い正答率はモデルパラメータ数と膨大な学習データ量によるものであると考える。LIama2 の正答率が低いことはモデルが比較的小さいことに加え、学習データのうち日本語割合が 0.1% と低い[14]ためと思われる。ChatGPT3.5 から ChatGPT4.0 への進化には僅か 1 年しか要しておらず、今後も LLM は急速に発展すると思われる。

現在の AI は既に画像を含めた臨床検査技師国家試験に合格する能力を保有している。すなわち国家資格としての臨床検査技師知識を持っていると言える。臨床検査技師として一人前になるためには国家資格を保有するのみでは不十分であり、On-the-Job Training として実務経験を積む必要がある。実務経験は患者接遇、採血、心エコー検査、腹部エコー検査、脳波検査、肺機能検査など患者さんを対象とした業務、血液や尿などの検体検査結果報告の為に解析手技、機器の操作方法、トラブルに対する対応、精度管理などの学習、病原体解析には染色、培養、遺伝子や蛋白質解析手技、病理では標本の作成や診断など多岐にわたる。そのため直ぐに臨床検査技師が AI に取って代わられることは無いと思われる。しかし他の医療従事者業務と同様に次第にその役割が AI よる置き換わりが進むことは必然であると考える。特に画像診断は AI が最も得意な分野であり、尿沈渣、末梢血液像（白血球分類）判定、細胞診判定、病理所見判定は最も早く AI 判定により置き換わると思われる。従来の検体検査、病理検査領域での臨床検査技師の役割は、AI を上手く使い、その誤りを正すことにシフトしていくと思われる。臨床検査技師の新たな業務領域としてタスク・シフト / シェアで実施可能になる業務や、体細胞や癌細胞の全遺伝子シークエンスを行い患者の体質や癌の性質に合わせた個別化医療法の提言などに広がる可能性がある。臨床検査技師教育においては、これまでの基礎的知識に加え、タスク・シフト / シェアで拡大した業務、AI に対する知識（原理、利用方法、限界）などが必要となると考える。今後も世の中は急速に変化し続けることから、常に勉強を続け新しいことを学習する態度を身につけることも求められる。

IV. 結　　論

今後の臨床検査技師教育においては、将来 AI を利用することを前提とし、AI の原理や活用法、ハルシネーションに代表される AI の限界などの教育も必要となる。

COI 状態

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) Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018; 18 (8): 500-10.
2) Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. Nat Rev Clin Oncol 2019; 16 (11): 703-15.
3) Funaita C, Furuie W, Koike F, Oyama S, Endo J, Otani Y, et al. Pattern recognition of forced oscillation technique measurement results using deep learning can identify asthmatic patients more accurately than setting reference ranges. Sci Rep 2023; 13 (1): 21608.
4) Yanagita Y, Yokokawa D, Uchida S, Tawara J, Ikusaka

M. Accuracy of ChatGPT on Medical Questions in the National Medical Licensing Examination in Japan: Evaluation Study. JMIR Form Res 2023; 7: e48023.

5) Taira K, Itaya T, Hanada A. Performance of the Large Language Model ChatGPT on the National Nurse Examinations in Japan: Evaluation Study. JMIR Nursing 2023; 6: e47305.

6) OpenAI. ChatGPT. available at: https://openai.com/chatgpt（accessed on 8th/April/2024）

7) Meta. Introducing Llama2. available at: https://ai.meta.com/llama/（accessed on 8th/April/2024）

8) Google. Bard. available at: https://bard.google.com/chat（accessed on 8th/April/2024）

9) 第 66 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 , 2020. available at: https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp200414-07.html（accessed on 8th/April/2024）

10) 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. available at: https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp210416-07.html（accessed on 8th/April/2024）

11) 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. available at: https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp220421-07.html（accessed on 8th/April/2024）

12) 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. available at: https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp230524-07.html（accessed on 8th/April/2024）

13) Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, et al. Emergent Abilities of Large Language Models. Transactions on Machine Learning Research（08/2022）available at: https://arxiv.org/abs/2206.07682（accessed on 8th/April/2024）

14) Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. GenAI, Meta. Llama 2: Open Foundation and Fine-Tuned Chat Models. available at: https://arxiv.org/abs/2307.09288（accessed on 8th/April/2024）