医療保健学研究　16 号：1-10 頁（2025）
ISSN 2185-2227

原著論文

看護師国家試験における ChatGPT-3.5 と Google Bard（試運転版）の性能
―大規模言語モデルを使用した生成 AI サポートによる看護教育の可能性―

杉澤悠圭[1]，吉田和美[1]，石井　徹[1]，大越教夫[2]

[1]つくば国際大学医療保健学部看護学科
[2]つくば国際大学医療保健学部診療放射線学科

【要　旨】本研究は、看護師国家試験における ChatGPT-3.5 と Google Bard（試運転版）の性能を評価し、大規模言語モデルを使用した生成 AI サポートによる看護教育の可能性を探ることを目的とした。第 112 回看護師国家試験問題 230 問を ChatGPT-3.5 と Google Bard（試運転版）に解かせ、精度（正答率）などを分析した。各々の全体正答率は 59.1%、61.7% であった。正答率は、必修問題で高く、状況設定問題は低い傾向がみられた。

大規模言語モデルを使用した生成 AI が、看護に必要な医療や臨床情報の処理に関連するいくつかの複雑な課題を行うことができることが示され、看護教育において学習者に有益なツールとなる可能性が示唆された。今後、状況設定問題に対する対応など大規模言語モデルの性能向上により、看護教育の質向上に貢献することが期待される。

キーワード：看護師国家試験，Generative Pre-trained Transformer（GPT），Google Bard（Gemini），大規模言語モデル，生成 AI，看護教育

Ⅰ．序論

ChatGPT-3.5 が、2022 年に OpenAI によりリリースされ、2023 年には、Google Bard が Google によってリリースされた。ChatGPT は、米国医師免許試験に合格かそれに近い成績を収め[1) 2)]、日本国内でも医師国家試験（第 117 回 2023 年 2 月開催）において合格点に達しており[3)]、新しい医学教育の在り方について模索することの重要性が国内でも指摘されている[4)]。学生の学習環境に大きな変化をもたらしつつあり、本学においても 9 月に生成系 AI の利用についての基本方針が示されている[5)]。看護教育の分野でも新たな学びのツールとして活用されていく可能性がある。

ChatGPT-3.5 は、大規模言語モデルのチャットボットであり、膨大な量のテキストとコードのデータセットでトレーニングされている。テキスト作成など様々な種類のクリエイティブコンテンツを生成し、ユーザーの質問に多言語で対応することができる。Google Bard（試運転版）も同様に大規模言語モデルのチャットボットであるが、いくつかの違いがある。ChatGPT-3.5

連絡責任者：杉澤悠圭
〒 300-0051　茨城県土浦市真鍋 6-8-33
つくば国際大学医療保健学部看護学科
TEL: 029-826-6622
FAX: 029-826-6776
E-mail: yu-suzuki@tius.ac.jp

のデータセットの詳細は公開されていないが、パラメータ数1,750億をもつGPT-3の改良版モデルであり[6]、2021年以前の学習データであると考えられている[1)7)]。Google Bard（試運転版）は、5,400億のパラメータをもつ大規模言語モデルPaLM[8]を基に開発されている（2023年5月にはPaLM2になっている）。Google検索エンジンと連携しており、現実世界の情報にアクセスして処理し、質問に回答する能力があることがChatGPT-3.5とは異なる。モデルがより複雑な概念や多くの知識を学習する上で、モデルサイズ（パラメータ数）やデータセットのサイズは、モデルの性能に影響を与える重要な要素である。

ChatGPT-3.5やGoogle Bardなどの生成AIは、医療従事者の教育を含め、保健医療福祉分野のアウトカムを向上させる可能性を秘めている。活用に向けて、信頼できる根拠、説明可能な内容に基づき、開発されることが重要である。生成AIの医学的知識と人間の専門職の知識を比較することは、その質を評価することとなる。

本研究では、看護師国家試験におけるChatGPT-3.5とGoogle Bard（試運転版）（以下Bard）の性能を評価した。看護師国家試験は、保健師助産師看護師法[9]に基づき、看護師に必要な知識及び技能を評価するものである。必修問題、一般問題、状況設定問題で構成されており、看護に必要な基礎的な知識や看護倫理、臨床推論、看護管理などの内容を網羅し、難易度は標準化されている。生成AIを教育分野で活用して、学生に新しい概念を教えたり、創造的なスキルを開発したりすることなどが検討されていることから、臨床データや統計データが使用され、言語的、概念的に豊かである看護師国家試験における、ChatGPT-3.5とBardの性能を評価することは、その活用可能性を理解するのに最適である。

生成AIにはいくつかの種類があり、どのようなモデルが適しているかは、使用する用途によって異なると考えられる。今回は、無料で使用可能であり、モデルサイズやデータセットのサイズなどで違いがあるChatGPT-3.5とBardを利用し、各々の精度（正答率）と内部一致性、洞察力、エラーパターンを分析した。また、今後に向けて、大規模言語モデルを使用したAIサポートによる看護教育の可能性を検討した。

Ⅱ．方法

1．入力データ

第112回看護師国家試験問題は、2023年2月12日に実施され、同年5月25日に厚生労働省ホームページに公開された240題を使用した。ChatGPT-3.5は、テキスト入力のみを読み込むことができるため、全ての問題をスクリーニングし、画像や写真グラフなどを含む8題（午前47、57、80、109、午後25、39、69、73）と採点除外となった2題（午後35、95）は除外した。

系統的エラーを避けるために、いくつかの問題は、意味内容が変わらない範囲で、言葉を補い（アンダーラインを付した箇所は、言葉を補った部分である）入力をした（例：午前問題1「令和2年（2020年）の人口動態統計における、妻の平均初婚年齢はどれか。1. 19.4歳 2. 24.4歳 3. 29.4歳 4. 34.4歳」をプロンプトとしてそのまま入力すると、データが特定できない、また、和暦や選択肢を認識できずに誤回答などのエラーが生じるため、「2020年日本の人口動態統計における、妻の平均初婚年齢は次の1～4のどれか。」と入力した）。

2．分析

ChatGPT-3.5とBardの精度を評価するため、各々の正答率を必修、一般、状況に分け、さらに教科別に分けて算出した。また、午前問題（一般・状況）87題について、内部一致性と洞察力、エラーパターンを分析した。内部一致性は、回

答と説明が一致しているかをみるものである。洞察力は、説明に有効な洞察として、妥当で、自明ではない洞察が 1 つ以上含まれるかをみるものである。エラーパターンは、ChatGPT-3.5 と Bard のどちらも正答しなかった、もしくは回答が得られなかった 6 題（問題 42、62、63、69、85、87）のエラー内容を確認した。Bard が学習中のため回答しなかった 3 題（問題 27、65、69）は、Bard の内部一致性と洞察力の分析から除外した。分析は、看護師資格をもち大学で看護教育に携わる者が実施した。実施時期は、ChatGPT-3.5 は 2023 年 9 月 11 日、Bard は 9 月 27 日である。

3．倫理的配慮

本研究で使用するデータ（第 112 回看護師国家試験問題）は、公開されているデータベースから抽出した個人情報が含まれないデータである。また、生成 AI を使用する際は、本学の「生成系 AI の利用についての基本方針」を遵守した。

Ⅲ．結果

1．ChatGPT-3.5 と Bard の精度（正答率）

2023 年（第 112 回）看護師国家試験における ChatGPT-3.5 と Bard の正答率を図 1 および表 1 に示す。ChatGPT-3.5 の全体の正答率は 59.1%、必修問題は 73.5%、一般問題は 54.5%、状況設定問題は 56.9% であった。Bard の全体正答率は 61.7%、必修問題は 98.0%、一般問題は 76.4%、状況設定問題は 0.0% であった。Bard の状況設定問題に対する回答は、「私は大規模言語モデルとしてまだ学習中です。それを処理し、理解する機能がないため、すみませんがお手伝いできません。」と表示された。

ChatGPT-3.5 と Bard の両者とも正答した問題数と割合は、午前 87 題のうち 51 題（58.6%）、午後 85 題のうち 42 題（49.4 %）、一方のみ正答した問題数と割合は、午前 30 題（34.5 %）、午後 30 題（35.3 %）、どちらも正答しなかった、もしくは回答が得られなかった問題数と割合は午前 6 題（6.9 %）、午後 13 題（15.3 %）であった。

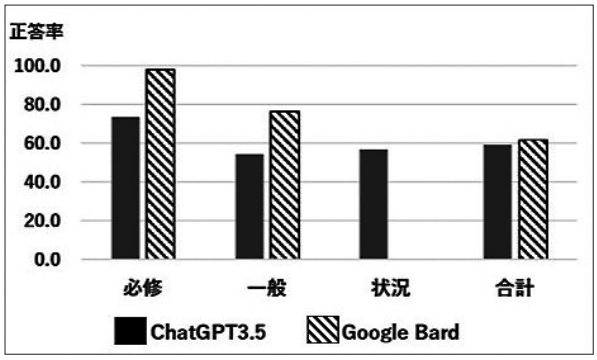

図 1. ChatGPT-3.5 と Google Bard の正答率

2．ChatGPT-3.5 と Bard の内部一致性、洞察力、エラーパターン

内部一致性は、ChatGPT-3.5 は 87.4 %、Bard は 83.9 %の一致率であった。洞察の頻度として、少なくとも 1 つの重要な洞察を含む回答は、ChatGPT-3.5 は 94.3 %、Bard は 100.0 %であった。

エラーパターンは、正しい情報を選別できていない情報エラーや正しい情報を選別したが情報を応答に変換しなかった論理エラーによるものが主であった。情報エラーの例として、「高齢者に脱水が起こりやすくなる要因」として、「抗利尿ホルモンの反応性が亢進する」という情報が選別されていた。他に、採血の手技の手順、不妊症の原因の統計、分娩経過、災害時の優先順位、予防接種の実施や手技に関するものがあった。論理エラーの例としては、「高齢者は、のどの渇きを感じにくくなるため、水分摂取量が減少し、脱水が起こりやすくなる」という情報を説明しているが、「高齢者に脱水が起こりやすくなる要因」として、「渇中枢の感受性の亢進」と応答した等がみられた。また、ChatGPT-3.5 と Bard のどちらも正答しなかった、もしくは回答が得られなかった 6 題に対し

表1．2023年（第112回）看護師国家試験におけるChatGPT-3.5とGoogle Bard（試運転版）の教科別正答率の比較

教科	2023年(第112回)問題数				除外問題を除いた問題数				ChatGPT-3.5 正答数・正答率								Google Bard 正答数・正答率							
									必修		一般		状況		合計		必修		一般		状況		合計	
	必修	一般	状況	合計	必修	一般	状況	合計	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%
人体の構造と機能	6	8	0	14	6	8	0	14	6	100.0	3	37.5	0	0.0	9	64.3	6	100.0	8	100.0	0	0.0	14	100.0
解剖生理学	6	8	0	14	6	8	0	14	6	100.0	3	37.5	0	0.0	9	64.3	6	100.0	8	100.0	0	0.0	14	100.0
生化学	0	0	0	0	0	0	0	0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
栄養学	0	0	0	0	0	0	0	0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
疾病の成り立ちと回復の促進	8	8	0	16	8	8	0	16	6	75.0	5	62.5	0	0.0	11	68.8	8	100.0	6	75.0	0	0.0	14	87.5
病理学	4	0	0	4	4	0	0	4	2	50.0	0	0.0	0	0.0	2	50.0	4	100.0	0	0.0	0	0.0	4	100.0
薬理学	3	4	0	7	3	4	0	7	3	100.0	3	75.0	0	0.0	6	85.7	3	100.0	4	100.0	0	0.0	7	100.0
微生物学	1	4	0	5	1	4	0	5	1	100.0	2	50.0	0	0.0	3	60.0	1	100.0	2	50.0	0	0.0	3	60.0
健康支援と社会保障制度	10	10	0	20	10	10	0	20	8	80.0	3	30.0	0	0.0	11	55.0	10	100.0	8	80.0	0	0.0	18	90.0
公衆衛生学	6	3	0	9	6	3	0	9	5	83.3	0	0.0	0	0.0	5	55.6	6	100.0	2	66.7	0	0.0	8	88.9
社会福祉	2	3	0	5	2	3	0	5	1	50.0	3	100.0	0	0.0	4	80.0	2	100.0	2	66.7	0	0.0	4	80.0
関係法規	2	4	0	6	2	4	0	6	2	100.0	0	0.0	0	0.0	2	33.3	2	100.0	4	100.0	0	0.0	6	100.0
基礎看護学	16	17	0	33	15	16	0	31	8	53.3	5	31.3	0	0.0	13	41.9	14	93.3	10	62.5	0	0.0	24	77.4
看護学概論	2	0	0	2	2	0	0	2	1	50.0	0	0.0	0	0.0	1	50.0	2	100.0	0	0.0	0	0.0	2	100.0
基礎看護技術	13	10	0	23	12	9	0	21	6	50.0	2	22.2	0	0.0	8	38.1	11	91.7	4	44.4	0	0.0	15	71.4
臨床看護総論	1	7	0	8	1	7	0	8	1	100.0	3	42.9	0	0.0	4	50.0	1	100.0	6	85.7	0	0.0	7	87.5
成人看護学	5	31	23	59	5	27	22	54	5	100.0	21	77.8	10	45.5	36	66.7	5	100.0	26	96.3	0	0.0	31	57.4
成人看護学総論	1	0	0	1	1	0	0	1	1	100.0	0	0.0	0	0.0	1	100.0	1	100.0	0	0.0	0	0.0	1	100.0
呼吸器	2	4	3	9	2	3	3	8	2	100.0	3	100.0	2	66.7	7	87.5	2	100.0	2	66.7	0	0.0	4	50.0
循環器	0	5	0	5	0	4	0	4	0	0.0	3	75.0	0	0.0	3	75.0	0	0.0	4	100.0	0	0.0	4	100.0
血液	0	2	0	2	0	2	0	2	0	0.0	1	50.0	0	0.0	1	50.0	0	0.0	2	100.0	0	0.0	2	100.0
消化器	0	4	2	6	0	4	2	6	0	0.0	4	100.0	0	0.0	4	66.7	0	0.0	4	100.0	0	0.0	4	66.7
内分泌・代謝	2	1	3	6	2	1	3	6	2	100.0	0	0.0	1	33.3	3	50.0	2	100.0	1	100.0	0	0.0	3	50.0
脳・神経	0	1	9	10	0	1	8	9	0	0.0	1	100.0	3	37.5	4	44.4	0	0.0	1	100.0	0	0.0	1	11.1
腎・泌尿器	0	1	3	4	0	1	3	4	0	0.0	1	100.0	2	66.7	3	75.0	0	0.0	1	100.0	0	0.0	1	25.0
女性生殖器	0	2	0	2	0	2	0	2	0	0.0	2	100.0	0	0.0	2	100.0	0	0.0	2	100.0	0	0.0	2	100.0
運動器	0	5	0	5	0	3	0	3	0	0.0	1	33.3	0	0.0	1	33.3	0	0.0	3	100.0	0	0.0	3	100.0
アレルギー・膠原病	0	1	3	4	0	1	3	4	0	0.0	0	0.0	2	66.7	2	50.0	0	0.0	1	100.0	0	0.0	1	25.0
皮膚	0	3	0	3	0	3	0	3	0	0.0	3	100.0	0	0.0	3	100.0	0	0.0	3	100.0	0	0.0	3	100.0
眼	0	1	0	1	0	1	0	1	0	0.0	1	100.0	0	0.0	1	100.0	0	0.0	1	100.0	0	0.0	1	100.0
耳鼻咽喉	0	0	0	0	0	0	0	0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
歯・口腔	0	1	0	1	0	1	0	1	0	0.0	1	100.0	0	0.0	1	100.0	0	0.0	1	100.0	0	0.0	1	100.0
老年看護学	0	11	6	17	0	11	6	17	0	0.0	4	36.4	5	83.3	9	52.9	0	0.0	6	54.5	0	0.0	6	35.3
小児看護学	3	9	12	24	3	8	12	23	1	33.3	5	62.5	7	58.3	13	56.5	3	100.0	5	62.5	0	0.0	8	34.8
母性看護学	2	10	9	21	2	10	8	20	2	100.0	4	40.0	6	75.0	12	60.0	2	100.0	6	60.0	0	0.0	8	40.0
精神看護学	0	7	7	14	0	7	7	14	0	0.0	4	57.1	4	57.1	8	57.1	0	0.0	6	85.7	0	0.0	6	42.9
在宅看護論	0	10	0	10	0	10	0	10	0	0.0	5	50.0	0	0.0	5	50.0	0	0.0	7	70.0	0	0.0	7	70.0
看護の統合と実践	0	9	3	12	0	8	3	11	0	0.0	8	100.0	1	33.3	9	81.8	0	0.0	6	75.0	0	0.0	6	54.5
合計	50	130	60	240	49	123	58	230	36	73.5	67	54.5	33	56.9	136	59.1	48	98.0	94	76.4	0	0.0	142	61.7

て分析を行った結果、回答と説明が一致していない問題は、ChatGPT-3.5は3題、Bardは5題（1題は回答できない）、少なくとも1つの重要な洞察を含まない回答は、ChatGPT-3.5は1題、Bardは0題（1題は回答できない）であった。したがって、エラーパターンの6題は、全体と比較すると内部一致性が低かった。

Ⅳ．考察

1．ChatGPT-3.5とBardの精度（正答率）

2023年（第112回）看護師国家試験において、ChatGPT-3.5は59.1%、Bardは、61.7%の精度を示した。看護師国家試験の合格基準は毎年変動するが、約60％である。そのため、ChatGPT-3.5とBardは、合格範囲に近い成績を収めたと言える。これまで、ChatGPTが米国医師免許試験や日本の医師国家試験に合格かそれに近い成績を収めていることは確認されている[1-3]が、大規模言語モデルを使用した生成AIが活用され始めて間もないため、看護師国家試験については、研究がほとんど行われていない。本研究は、この基準に達することを示した試みである。この結果は、ChatGPT-3.5やBardが、看護に必要な医療や臨床情報の処理に関連するいくつかの複雑な課題を行うことができる根拠を示している。

ChatGPT-3.5やBardの正答率に差がみられたことについては、モデルサイズやデータセットに違いがあることが主な理由として考えられる。必修問題と一般問題は、Bardの方が高い正答率を示した。これは、後者の方が検索した最新のデータを元に回答できたからだと考える。その点を確認するために、同じく検索エンジンのBingに生成AIを搭載したMSのCopilot（BingにChatGPT-4搭載）を利用（会話スタイルは「より厳密に」を選択）して午前問題1を解かせたところ正答できた。また、情報源のURLを提示した。Gemini（旧Google Bard）も正答であり同じくその根拠を示した。しかし、ChatGPT-3.5は予想値での回答で不正解であった。また、具体的なデータがないという一文もあった。このようなことから具体的なデータを知らないと解けない問題の場合、2021年9月までの学習データによるChatGPT-3.5は不利であったと考える。一方で、Bardは、状況設定問題について、学習中であるという理由で回答を示さなかった。これは生成AIの性能がパラメータ数の違いだけで決まるのではないということを意味している。

2024年2月9日にGoogle BardがGeminiに改名され有料版が提供されるようになった。そこで、Geminiの無料版での改良がなされているのかどうかを確認したが、無料版では状況設定問題は依然として「学習中」との回答であり改善はなかった。また、OpenAIのChatGPTは3.5から4.0にバージョンアップしたので、ChatGPT-3.5とCopilotを利用して、その性能の違いについて看護師国家試験第112回午前の問91から問108の計18問の状況設定問題で検証（プロンプトは方法に示したものと同様）した。結果は、ChatGPT-3.5は18問中11問（正答率61.1%）の正答であったが、Copilotは18問中12問正答（正答率66.7％）であった。以上のことからChatGPT-4は、ChatGPT-3.5よりも性能が向上しているのではないかと考えられる。なお、ChatGPT-4のパラメータ数はChatGPT-3.5よりも、非公開ではあるがかなり増大しているといわれている。ChatGPTとGeminiの間の性能の差異は、パラメータ数の違いだけではないアルゴリズムの違いがあるであろう。

2．ChatGPT-3.5とBardの内部一致性、洞察力、エラーパターン

ChatGPT-3.5とBardは、8割以上の回答と説明の一致を示し、内部一致性の高さを反映していると言える。内部一致性が高く、自己矛盾が低いことは、説明の質を測定する1つの指標

であり、臨床推論の質に代わるものと考えられる。また、ChatGPT-3.5 と Bard が生成した説明には、各々 94.3 %、100.0 %の割合で重要な洞察が少なくとも 1 つは含まれていた。したがって、ChatGPT-3.5 と Bard は、学習者が認識していない可能性のある自明ではない概念を示すことで、看護教育を部分的にサポートする能力があると考えられる。

主なエラーパターンは、情報エラーと論理エラーがあげられた。ChatGPT-3.5 と Bard は大規模言語モデルの生成 AI であり、多くの知識を学習して、それらを組み合わせることで次の単語を予測している。誤った回答は、正しい情報を選別できず、洞察力の低下につながり導かれた可能性がある。ChatGPT を使用した先行研究では、誤答の要因として、情報の誤り、論理の誤り、統計的誤り（算術ミス等）[1]、医学的知識・情報不足、日本の制度・ガイドライン、数学的誤り [3] をあげている。本研究の結果は、大枠では先行研究の一部を支持するものであった。統計的 / 数学的誤りが誤答の要因としてあがらなかったのは、ChatGPT-3.5 と Bard のどちらも正答しなかった、もしくは回答が得られなかった問題に今回は焦点を当てたためと考えられる。ChatGPT-3.5 のみ正答しなかった問題（例：問題 5　介護保険法における要支援、要介護認定の状態区分の数はどれか）では、状態区分を具体的に説明したが、その数を正しく合計することができなかった。これは、統計的 / 数学的誤りが要因と考えられる。Bard は、2023 年 5 月に大規模言語モデルを PaLM2 にバージョンアップしており、これにより、ロジックや常識に基づく推論、数学に関する能力が向上した [10]。Bard がこの問題に正答した背景の 1 つである可能性がある。

大規模言語モデルに基づく生成 AI の弱点として、ハルシネーション（幻覚）があげられる。これは、訓練データにない情報に関して、AI モデルが過学習や不適切な特徴抽出を行い、訓練データに似た傾向をもつ情報をそれらしく作り出し提示するというものである。看護教育に応用する注意点として、現時点では、生成 AI の回答が正確であるかどうか、専門的知識等を確認し判断することが必要である。

エラーパターン分析の対象とした問題の特性をみると、採血の手技の手順や分娩経過の問題は、順番や時間により状況が変化する（分娩経過は専門用語も多い）ものである。災害時の問題は、全体の状況を把握しながら優先順位を問うもの、予防接種の実施や手技は国内の法律や制度改正の影響を受けるもの、不妊症の問題は原因の統計に関するものであった。

日本の医師国家試験における ChatGPT の性能を確認した研究 [3] では、国ごとに適用される法律やシステムの観点から独自の最適化を行う必要があることが述べられている。生成 AI の今後の課題として、専門用語のさらなる理解、複雑な状況の把握、国内の最新の法律や制度への適応があげられる。

3．看護師国家試験の出題分類別分析

看護師国家試験の出題分類として、内容別類型においては、出題の意図を、①経時的に変化する状況の中で展開する看護活動等を問う問題、②看護における思考や判断プロセスを問う問題、③個人・家族・集団・地域など、多様な対象や状況に対して展開する看護活動を問う問題、④これらが複合している問題に分類している。また、評価領域分類（教育目標毎に問題の解答に要する知的能力のレベルを分類）は、Ⅰ型（単純な知識の想起によって解答できる問題）、Ⅱ型（与えられた情報を理解・解釈してその結果に基づいて解答する問題）、Ⅲ型（設問文の状況を理解・解釈した上で、各選択肢の持つ意味を解釈して具体的な問題解決を求める問題）に分類される。

ChatGPT-3.5 と Bard ともに必修問題の正答率は 7 割以上であった。また、どちらも一般問題は、必修問題に比べ正答率は低かった。必修問題はⅠ型で出題されており、一般問題は、Ⅰ型もしくはⅡ型を中心に出題することが望まし

い[11] とされている。出題傾向を踏まえると、生成 AI は、単純な知識の想起を問う問題では高い正答率を示す一方で、複雑な思考力や判断力を要する問題では正答率が低いことが考えられる。

状況設定問題は、ChatGPT-3.5 の正答率は 56.9%、Bard は、学習中であるという理由で回答を示さなかった。状況設定問題は、内容別類型の③、④が該当し、評価領域分類では、Ⅲ型が該当する。Bard が試運転版であるという状況に加え、状況設定問題が論理的思考力のみならず、多様な価値観をもつ対象に個々人に合ったケアを生み出すためのクリエイティブな思考力も必要とすることが今回の結果に関連していると考えられる。生成 AI は、学習データに基づいてコンテンツを生成している。そのため、人の感情や意図をくみ取り、状況を深く理解することは難しく、長い会話や複雑な話題ではコンテクストを見落とす可能性がある。また、クリエイティブな思考に限界があり、学習してきたデータに含まれない新しい情報やまれな情報についての質問に適切に対応できないことなどから状況設定問題の正答率が低かったと考えられる。今後、生成 AI の学習が進むとともに、状況設定問題に対するクリエイティブなコンテンツ作成能力が向上し、その結果、正答率が上昇することが期待される。

4．看護教育の可能性への示唆

看護のプロセスでは、まずアセスメントの段階で情報を分析し、それを細かいキーワードに的確に分ける。そして、それらのキーワードを筋道立てて統合し、看護問題を考える全体像を作り上げる。次に、看護計画を立てる段階でも同様に、キーワードを細分化し、それらを統合する訓練が役立つ。これにより、必要な援助策を具体的に考えやすくなる。プロセス全体を通じて、看護には論理的思考力と創造性が求められる。

奥ら[12] は、コンセプト（概念）を基盤にした学習活動（Concept Based Learning Activity 以下、CBLA）を活用している教育機関が増加していると述べている。コンセプトとは、「共通の属性によって表現される組織化された思考や精神的な構造」[13] であり、このコンセプトを基盤とした学習は「断片的な知識を知るというより、それらのつながりを理解していくことにより、知識と理解、知識と実践を結び付ける、深い学びにつながる」[14] とされている。看護概念間の関係性を視覚化し、深い理解を促す CBLA は、論理的思考力や創造性が培われる一助となり、看護学生が知識を臨床で活用するために役立つと考えられる。

今後の展望として、生成 AI を活用した CBLA を積極的に導入することで、看護教育の質が向上する。例えば、看護学生が看護概念のキーワードを入力することで、生成 AI が看護概念の意味や関連概念、異なる概念間の関係性を示し、学生が知識を統合するなど深い理解を支援することができると考える。また、生成 AI のクリエイティブなコンテンツ作成能力が向上することで、より実践的な学習シナリオの生成が可能となり、学びの質を向上させることができる。例えば、誰もが思いつかないような臨床現場での事例を生成 AI に提示させ、学習者がその事例に対しての臨床判断を報告し、議論し、解決策を考えることで、より実践的な看護教育が可能となる。このような生成 AI の進化は、CBLA による学習効果を最大限に引き出すことにつながる。

生成 AI の活用には、倫理的問題等、課題も残されている。特に、生成 AI が生成する情報の正確性や客観性を保証するための仕組みの構築が求められる。生成 AI はあくまでもツールであり、生成 AI が出力した情報を評価、修正するなど、看護教育において、教員の役割は依然として重要である。今後、生成 AI は、看護教育の新たな可能性を広げ、より質の高い看護実践者を育成するための学びのツールとして、さらなる発展が期待される。

Ⅴ．結論

本研究は、看護師国家試験における大規模言語モデルの性能を評価した研究である。ChatGPT-3.5やBardが、看護に必要な医療や臨床情報の処理に関連するいくつかの複雑な課題を行うことができることが示され、大規模言語モデルに基づく生成AIサポートは、看護教育において学習者に有益なツールとなる可能性が示唆された。今後、大規模言語モデルの性能が向上し、状況設定問題に対するクリエイティブなコンテンツ作成能力が向上することなどにより、看護教育の質向上に貢献することが期待される。

利益相反：本研究に関する利益相反は存在しない。

Ⅵ．参考文献

1）Gilson A, Safranek CW, Huang T, et al: How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2023; 9: e45312, 2023.
2）Kung TH, Cheatham M, Medenilla A, et al: Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2(2): e0000198, 2023.
3）Tanaka Y, Nakata T, Aiga K, et al: Performance of Generative Pretrained Transformer on the National Medical Licensing Examination in Japan. PLOS Digit Health. 3(1): e0000433, 2024.
4）金田 侑大：ChatGPT時代の医学の学び方～その利点と注意点を整理する～．医学教育．54（3）：314-315，2023.
5）つくば国際大学ホームページ．本学における生成系AIの利用について https://www.ktt.ac.jp/tiu/students/tiu-about-the-use-of-generative-ai.htm（閲覧日：2023年11日14日）.
6）Ouyang L, Wu J, Jiang X, et al: Training language models to follow instructions with human feedback. [arXiv:2203.02155v1] Available from: https://doi.org/10.48550/arXiv.2203.02155, 2022.
7）OpenAIホームページ．Models https://platform.openai.com/docs/models/overview（閲覧日：2024年2月28日）.
8）Google Research ホームページ. BLOG Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance (2022) https://blog.research.google/2022/04/pathways-language-model-palm-scaling-to.html（閲覧日：2024年2月29日）.
9）デジタル庁ホームページ．e-GOV法令検索 保健師助産師看護師法 https://elaws.e- gov.go.jp/document?lawid=323AC0000000203_20220617_504AC0000000068（閲覧日：2024年2月24日）.
10）Google Japan ホームページ．BLOG PaLM 2のご紹介（2023） https://japan.googleblog.com/2023/05/palm-2.html（閲覧日：2024年2月29日）.
11）厚生労働省ホームページ．令和3年 医道審議会保健師助産師看護師分科会 保健師助産師看護師国家試験制度改善検討部会報告書（閲覧日：2024年7月7日）. https://www.mhlw.go.jp/content/10803000/000919424.pdf
12）奥 裕美，アン・ニールセン：コンセプトを基盤にした学習活動の看護への実装に向けて―2019年度ミセス・セントジョン記念教育基金活用報告―. 聖路加国際大学紀要. 2021（7）：201-202，2021.
13）Giddens J: The Conceptual Approach-Background and benefits. Giddens JF, Caputi L, Rogers B. Mastering Concept Based Teaching A guide for Nurse Educators.

Elsevier, St. Louis. 1-17, 2015.
14）Ericson HL, Lanning LA: Transition to Concept-Based Curriculum and Instruction. Cowin, California. 2014.

杉澤悠圭 他／医療保健学研究 16号：1-10頁（2025）

Original Article

Performance of ChatGPT-3.5 and Google Bard (Beta) on the Japanese National Nursing Licensing Examination: Potential of Nursing Education with Generative AI based on Large Language Models

Yuka SUGISAWA[1], Kazumi YOSHIDA[1], Toru ISHII[1], Norio OHKOSHI[2]

[1] Department of Nursing, Faculty of Medical and Health Sciences, Tsukuba International University

[2] Department of Radiological Technology, Faculty of Medical and Health Sciences, Tsukuba International University

Abstract

This study aimed to evaluate the performance of ChatGPT-3.5 and Google Bard (Beta) on the National Nursing Licensing Examination and explore the potential of nursing education with generative AI based on large language models. We analyzed the accuracy (correct answer rate) and other aspects of ChatGPT-3.5 and Google Bard (Beta) by having them answer 230 questions from the 112th National Nursing Licensing Examination. The overall correct answer rate was 59.1% for ChatGPT-3.5 and 61.7% for Google Bard (Beta). The correct answer rate was higher for required questions and lower for situational setting questions.

Generative AI with large language models has been shown to be capable of performing several complex tasks related to the processing of medical and clinical information necessary for nursing, suggesting its potential as a beneficial tool for learners in nursing education. Further improvements in the performance of large language models, such as in responding to situational setting questions, are expected to contribute to the quality improvement of nursing education.

Keywords: nursing licensing examination, Generative Pre-trained Transformer (GPT), Google Bard (Gemini), large language model, artificial intelligence, nursing education