News Release

金沢大学
KANAZAWA UNIVERSITY

令和６年３月１８日

各報道機関文教担当記者　　殿

最適化されたプロンプトとGPT-4により

ChatGPTが日本の医師国家試験に合格可能な成績を達成！

金沢大学融合研究域融合科学系の野村章洋准教授と株式会社 MICIN の共同研究グループは，**ChatGPT に日本の医師国家試験を解かせるために最適化されたプロンプト（※1）を開発しました。さらに，このプロンプトと GPT-4 を用いることで、最低合格得点率を上回ることに成功しました。**

2023 年初頭，ChatGPT にアメリカの医師国家試験（USMLE）を解かせたという論文がジャーナルに公開されて以降，ChatGPT の医療・ヘルスケア分野での活用可能性が世界的に大きな注目を浴びるようになりましたが，英語圏以外の医師国家試験での研究は発展途上でした。

本研究では，まず第 116 回医師国家試験（2022 年 2 月実施）の問題の中から画像データを有さない 290 問を基に，GPT-3.5 ならびに GPT-4 を用いて最も正答率の高いプロンプトを決定しました。次に，その最適化されたプロンプトを用いて GPT-4 モデルを搭載した ChatGPT に，第 117 回医師国家試験（2023 年 2 月実施）を解かせたところ，必修問題で 82.7%，基礎・臨床問題で 77.2%のスコアを獲得し，それぞれ最低合格得点率を上回る結果となりました。

さらに，ChatGPT が誤答を出力した原因の詳細分析を行いました。その結果，医学知識の不足や日本特有の医療制度に関する情報不足，計算問題での誤りなどが誤答要因であることが分かりました。

本研究結果より，実際の医療現場での運用にはまだ課題が残りますが，**ChatGPTが日本の医師国家試験の最低合格得点を超える可能性を持つことが示されました。また，近い将来，大規模言語モデルが日本国内の医療現場において活用される医療用AIの基盤モデルの一つとなることが期待されます**。

本研究成果は，2024 年 1 月 23 日に国際学術誌『*PLOS Digital Health*』にオンライン掲載されました。

【研究の背景】

GPT（Generative Pretrained Transformer，※2）に代表される大規模言語モデル（LLM，※3）の登場により，臨床現場や医学研究の支援ツールとしての応用が期待されています。2023年初頭に，ChatGPTが米国の医師国家試験（USMLE）で合格点に迫る性能を示し話題となりましたが，英語以外の言語圏での医師国家試験における性能は，これまで十分に評価されていませんでした。

【研究成果の概要】

本研究では，GPT-3.5ならびにGPT-4モデルが搭載されたChatGPTを用いて，ChatGPTへの指示（プロンプト）を最適化した上で，日本の医師国家試験を解答させた際の正答率を検討しました。まず，2022年2月に実施された第116回医師国家試験を用いて，問題内に画像データを有さない290問を，プロンプト最適化用データセットとして使用しました。そして正答率が最も高くなるようなプロンプトを決定しました。その最適化されたプロンプトを用いて，GPT-4モデルが搭載されたChatGPTに入力し，2023年2月に行われた第117回医師国家試験から問題内に画像データを有さない262問を用いてその正答率を評価しました。その結果，必修問題で82.7%，一般問題で77.2%の正答率を達成しました。，これは，同回試験における受験生の最低合格得点率をいずれも上回っていました。さらに，ChatGPTが誤答を出力した原因を分析したところ，医学知識の不足，日本特有の医療制度に関する情報不足，そして計算問題での誤りの3点が主な要因だったことが分かりました。

【今後の展望】

本研究により，GPT-4と最適化されたプロンプトを共に用いることで，ChatGPTは日本の医師国家試験で最低合格得点を超える可能性があることが示されました。このような LLM は，人間の受験者を対象とした問題に解答するという次元を超えて，医療・ヘルスケア分野のアンメット・メディカルニーズを満たす最良の「相棒」となる可能性を秘めていると考えられます。ただし，現時点では，専門的な医学知識をどのように継続的に学習をさせ，さらに最新かつ正確な出力を担保するかなどの課題が残っています。しかしながら，このような LLM は近い将来，臨床的有効性と安全性が科学的に示され，医療用AIの基盤モデルの一つとなることが期待されます。

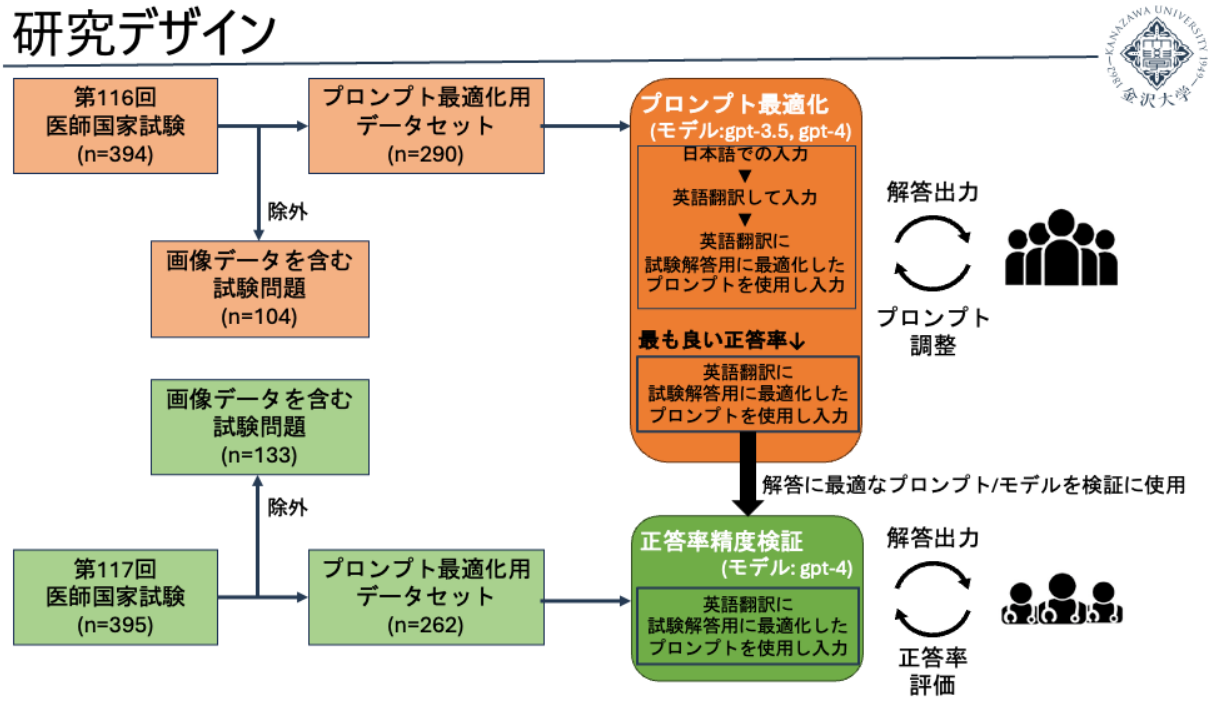
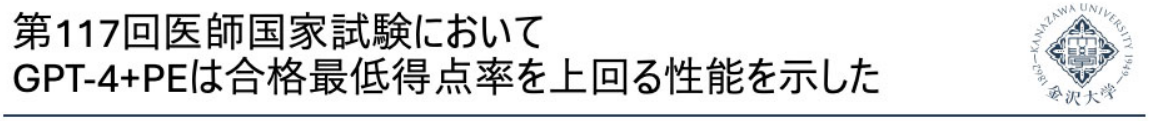


図 1. 研究デザイン



	必修問題			一般・臨床問題				
	一般問題	臨床問題	長文問題	一般問題 (総論)	一般問題 (各論)	臨床問題 (総論)	臨床問題 (各論)	長文問題
問題数	45	22	15	61	27	36	46	10
正答数	36	19	12	47	25	22	37	8
出力エラー表示数	1	0	0	0	1	0	0	0
不正答数	8	3	3	14	1	14	9	2
正答率	80.0%	86.4%	80.0%	77.0%	92.6%	61.1%	80.4%	80.0%
出力エラー表示率	2.2%	0.0%	0.0%	0.0%	3.7%	0.0%	0.0%	0.0%
点数の重み付け	x1	x3				x1		
総合得点 (総合正答率)	129/156 (82.7%)					139/180 (78.3%)		
合格最低ライン	80.0%					74.6%		

図 2. GPT-4 と最適化プロンプトによる第 117 回医師国家試験の得点と総合正答率

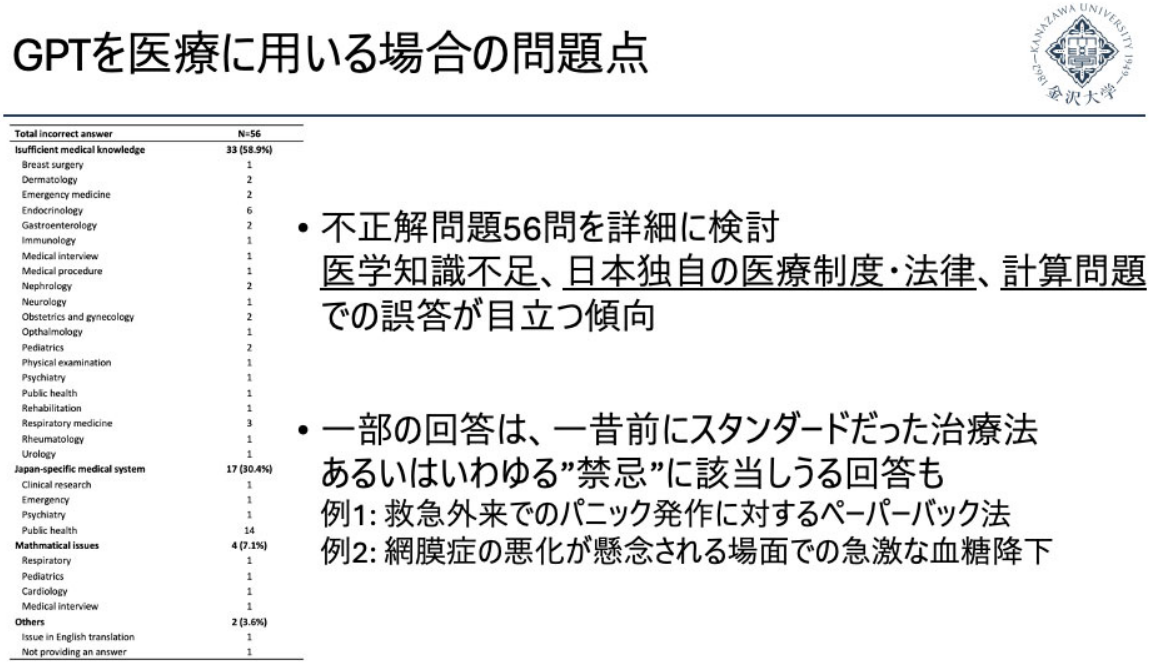


図 3. 誤答出力の検討と医療に用いる場合の問題点

【掲載論文】

雑誌名：*PLOS Digital Health*

論文名：Performance of Generative Pretrained Transformer on the National Medical Licensing Examination in Japan.
（日本医師国家試験を用いた大規模言語モデル（GPT）の解答性能の評価）

著者名：Yudai Tanaka, Takuto Nakata, Ko Aiga, Takahide Etani, Ryota Muramatsu, Shun Katagiri, Hiroyuki Kawai, Fumiya Higashino, Masahiro Enomoto, Masao Noda, Mitsuhiro Kometani, Masayuki Takamura, Takashi Yoneda, Hiroaki Kakizaki, Akihiro Nomura

（田中雄大，中田卓人，相賀耕，惠谷隆英，村松諒太，片桐駿，河合弘行，東野史弥，榎本真大，野田昌生，米谷充弘，高村雅之，米田隆，砥﨑裕晃，野村章洋）

掲載日時：2024 年 1 月 23 日にオンライン版に掲載

DOI：10.1371/journal.pdig.0000433

【用語解説】

※1　プロンプト

人間が、対話型生成 AI に入力する指示内容のこと。プロンプトの内容を工夫することで、大規模言語モデルはそのままに、より人間側が意図するタスクをモデルに行ってもらい、その性能を向上させることが可能となる場合がある。

※2　GPT （Generative Pretraind Transformer）

OpenAI 社が開発した Transformer と呼ばれる機械学習アルゴリズムをベースに改良が加えられた言語モデルのこと。

※3　大規模言語モデル（Large Language Model, LLM）

自然言語処理を行う言語モデルの学習を、大量のデータを用いて大規模に行ったもの。なお、自然言語処理とは、人間が扱う言語（自然言語）を対象に, その理解や生成を計算機に処理させる技術の総称。ある文章に続く単語の生成確率を、機械学習を経て算出する言語モデルを用いることが多い。

【本件に関するお問い合わせ先】

■研究内容に関すること

金沢大学融合研究域融合科学系　准教授

野村 章洋（のむら あきひろ）

TEL：076-265-2259

E-mail：anomura@med.kanazawa-u.ac.jp

■広報担当

金沢大学融合系事務部総務課企画総務係

荒井　創（あらい　つくる）

TEL：076-264-5920

E-mail：yugosomu@adm.kanazawa-u.ac.jp