厚生労働科学研究費補助金（がん対策推進総合研究事業）
（分担研究報告書）

参照する情報源を限定した生成AIチャットボットによるがん情報提供のハルシネーション排除の可能性

研究協力者　西迫 宗大　国立がん研究センター がん対策研究所 がん情報提供部（特任研究員）
研究分担者　東 尚弘　　東京大学大学院 医学系研究科 公衆衛生学分野（教授）
研究代表者　若尾 文彦　国立がん研究センター がん対策情報センター本部（副本部長）

研究要旨

本研究では、人工知能 (AI) による自然言語処理技術を備えたチャットボットを用いた信頼性の高いがん情報の提供の可能性について検討した。確かな情報を参照データとした生成AIチャットボットの返答の特徴を把握する事を目的とし、参照する情報源が異なる生成AIチャットボットを試作・比較した。参照する情報源は①ChatGPT（Generative Pre-trained Transformer, OpenAI)自体（参照先を指定しない）に対して ②国立がん研究センターが運営する「がん情報サービス」③Google検索エンジンにより上位に掲出されたWebサイトの3種類とした。これらにがんに関する質問に対する返答を生成させた。

その結果、①は全ての質問に対して返答したが、ハルシネーションが1割程度みられた。対して②は、全てのがん情報サービスに存在しない情報の質問に対して「情報が存在しないため返答できない」と答え、ハルシネーションは見られなかった。しかし、がん情報サービスに存在する情報の質問に対して返答しないパターンも見られた。③は②と比較し返答しないパターンは少ないが、その返答にはハルシネーションを 1 割程認め、科学的エビデンスのない情報に対して真実のように断言する表現もみられた。

確かながんに関する情報群を参照情報として限定した生成AIチャットボットは、ハルシネーションを減少できる利点を認めた。しかし、科学的エビデンスのない情報に対してその説明や理由を返答することはなく、結果的に返答できる範囲が限定された。これらの解消には、参照する情報源となるWebサイト自体のテキストの工夫や質問内容に対するプロンプトの改良が必要と考えられた。

A．研究目的

発達したインターネットやソーシャルメディアは、患者が情報を得たり意見を理解したりする機会を提供し、治療を決定するのに役立っている。その一方で、Web上に溢れる誤った医療情報が問題視されている[1,2]。がん領域において、長大な情報の中からどの情報源が正しく、エビデンスに基づいた科学的に妥当な情報なのかを患者が判断するのは困難であり、誤った情報による、適切な医療を受ける機会の損失、副作用、医療費に関する経済的損失が問題となっている[3]。

人工知能（Artificial Intelligence, AI）の発達は、コンピューターが人間に代わって会話をする自動会話プログラムの進歩にも貢献を果たした。医療情報検索において、AIによる自然言語処理を用いた生成AIチャットボットは、従来の検索エンジンと比較して有利な点が報告されている[4-6]。がんに関する情報の情報の検索において、生成AIチャットボットの使用により一般の健康リテラシーが向上し、簡単にアクセスできる理解しやすい健康情報が入手できる可能性がある。

しかし、生成 AI チャットボットを用いた情報検索において、誤った情報を含む返答（ハルシネーション）の問題は解決されていない[5,6]。現在一般に広く使用されている ChatGPT が回答を生成するための学習データは世界中のインターネット上に公開されている大量のテキストデータとされており[5]、回答を生成するために使用した参照するデータを確認する事ができない[6]。

そこで我々は、参照する情報を限定した生成 AI チャットボットを試作し、人工知能を用いた信頼性の高いがん情報の提供の可能性を検討した。本研究では、確かな情報を参照データとした生成 AI チャットボットの返答の特徴を把握する事を目的とした。

B. 研究方法

参照する情報源が異なる生成AIチャットボットを試作・比較した。参照する情報源は①ChatGPT (ChatGPT, Generative Pre-trained Transformer, OpenAI) 自体（参照先を指定しない）に対して ②国立がん研究センターが運営する「がん情報サービス（https://ganjoho.jp）」)； ③Google検索エンジンにより上位に掲出されたWebサイトの3種類とした。①は通常のChatGPT-3.5-turboを用いて回答を生成させた（図1- A）。②は「がん情報サービス」のテキストをベクトル化しナレッジデータベースに収容した。この中より質問事項に対する類似性の高いテキストをベクトル検索することにより質問に関する情報を得た。OpenAI社から提供されているアプリケーション・プログラミング・インターフェースを活用し、得られたテキストをプロンプトエンジニアリングにより大規模言語モデル（LLM）へと質問内容を渡し、回答を生成させた(図1-B, 図2-A)。③は、Web検索ツールであるSerpAPI（LangChain）を連携させ、LLMがGoogle検索により情報を得て回答を生成するようにシステムを構築した(図1-C, 図2-B)。②,③のLLMはGPT-3.5-turboおよびGPT-4.0とした。本システムの環境構築は生成型人工知能開発企業（株式会社pipon, 東京都）に依頼し参照する情報源を限定した上記の生成AIチャットボット環境が構築された。

がんに関する質問事項として、がん情報サービス内に存在する事項/しない事項についてそれぞれ20問が作成された。これらを含むプロンプトに対し①～③により返答されたテキストについて、

1. 返答の有無
2. 返答しない場合の追加情報の有無
3. 曖昧な返答の有無
4. ハルシネーションを含む返答の有無

を確認し集計した。

（倫理面への配慮）

本研究は、個人情報を取り扱うことはない。したがって、個人情報保護上は特に問題は発生しないと考える。

C. 研究結果

1. 返答の有無（表１）

構築された生成AIチャットボットごとの質問に対する返答の有無について、全体の集計では①は40問すべての質問に対して返答した。②はGPT-4をモデルとした場合の40問の質問に対する返答割合は35.0%であり、同様にモデルがGPT-3.5の場合は47.5%であった。③はGPT-4, 57.5%; GPT-3.5, 85.0%であった。がん情報サービスに存在する内容20問での質問に対し②はモデルGPT-4では、70.0%, GPT-3.5; 95.0%の返答割合であった。同様に、③はGPT-4, 55.0%; GPT-3.5; 95.0%であった。がん情報サービスに存在しない内容での20問の質問に対し②はすべて返答しなかった。③はモデルGPT-4では、60.0%, GPT-3.5; 75.0%の返答割合であった。

2. 返答しない場合の追加情報の有無（表2）

1.の結果において、返答しない場合、追加の情報を与えたかを見た。全体の集計では②はGPT-4をモデルとした場合の40問の質問に対する追加情報を与えた割合は15.4%であり、同様にモデルがGPT-3.5の場合は52.4%であった。③はGPT-4, なし; GPT-3.5, 16.7%であった。がん情報サービスに存在する内容での20問の質問に対し②はモデルGPT-4では、16.7%, GPT-3.5; 100%の割合であった。同様に、③はGPT-4, -3.5とも追加の情報を返答しなかった。がん情報サービスに存在しない内容での20問の質問に対し②はGPT-4, 15.0%; GPT-3.5; 50.0%であった。③はモデルGPT-4追加情報なし, GPT-3.5; 20.0%の割合であった。

3. 曖昧な返答の有無（表3）

生成AIチャットボットが生成した返答において、曖昧な内容の有無について集計した。①は40問の質問に対して10.0%の返答でみられた。②はGPT-4, -3.5モデルともに曖昧な返答は見られなかった。③はGPT-4, 21.7%; GPT-3.5, 25.7%であった。がん情報サービスに存在する内容での20問の質問に対し①は5.0%、③GPT-4, 18.2%; GPT-3.5, 15.8%であった。がん情報サービスに存在しない内容での20問の質問に対し①は15.0%、③はモデルGPT-4では25.0%, GPT-3.5; 37.5%の割合であった。

4. ハルシネーションを含む返答の有無（表4）

生成AIチャットボットが生成した回答におけるハルシネーションの出現を集計した。全体の集計では①は40問の質問に対して7.5%認めた。②はハルシネーションを含む返答は認めなかった。③GPT-4, 13.0%; GPT-3.5, 11.8%であった。がん情報サービスに存在する内容での20問の質問に対し①は5.0%、③はGPT-4, -3.5ともにハルシネーションを含む返答は認めなかった。がん情報サービスに存在しない内容での20問の質問に対し①は10.0%、③はモデルGPT-4では25.0%, GPT-3.5; 26.7%の割合であった。

D. 考察

参照する情報源を限定した生成AIチャットボットのがんに関する質問の返答は、その参照する情報源により様々な特徴を示した。正しいがんの情報を得る方法として、確かな情報（＝「がん情報サービス」）を参照情報とした生成AIチャットボットを開発する事は、長大ながんに関する情報の中からエビデンスに基づいた科学的に妥当な情報を得ることのできる可能性を持つが、現状では様々な改良が必要なことも明らかとなった。

一般に公開されているChatGPTは、がんに関する40問の質問に対しすべて返答した事に対し、参照情報を限定した生成AIチャットボットでは「情報が存在しないため返答できない」と答え、質問に対して返答しない事象が多くみられた。その傾向は、がん情報サービスを情報源とした場合と、モデルがGPT-4の場合に多く見られた。がん情報サービスを情報源とした生成AIチャットボットの場合、がん情報サービスに存在しない情報についての質問に対して返答する事はなかったが、逆にがん情報サービス内に存在する情報に対して回答を生成しない場合も認めた（GPT-4, 30.0%; -3.5, 5.0%）。この場合、データベース上に存在する情報はエビデンスに基づいた情報のみとなり、逆にエビデンスに基づく事の無い情報は参照する事ができない。結果として、不確かな情報を「不確か」とは返答できず、あらゆる・一般的な質問に対する回答範囲が狭くなったと考えられた。また、LLMによる返答有無割合の差の理由に関し、詳細は不明だが、モデルの性能向上による、参照情報の必要量の差異によるものが影響した可能性が考えられた。

がんに関する質問に対し「情報が存在しないため返答できない」と返答したうえで、何らかの情報を与えた事象は、参照情報源をがん情報サービスとした生成AIチャットボットがGoogle検索を情報源とした生成AIチャットボットよりも多く認めた。多くは「医師や専門家に相談する事」を追加の情報として与えたが、中には「（がん治療の相談先として）がん相談支援センターやがん診療連携拠点病院の相談員にも相談することができます。これらの機関は、あなたと担当医の橋渡しをしてくれることできます。」のような有益な情報を与えた返答も認めた。

エビデンスが確立されていない事象に対する返答において「～の可能性がある」「～と言われている」のように曖昧に返答したパターンがみられた。その割合がもっと多い傾向にあったのは、Google検索を参照情報源とした生成AIチャットボットであり、特に「がん情報サービス」に存在しない情報に対する回答で多く見られた。一般公開されているChatGPT-3.5の返答も40問の質問に対して、10%で曖昧な表現が含まれる回答が生成された。質問事項が、がん情報サービス内に含まれていたか、いなかったかで見ると、がん情報サービスに存在しない情報に関する質問に対する返答の方が、存在する情報に対する返答に対して曖昧な返答が多く含まれる傾向がみられた。一方、がん情報サービスを参照情報とした生成AIチャットボットの回答では曖昧な返答は見られなかった。がん情報サービスには、エビデンスに基づく情報が掲載されており、がん情報サービスを参照情報とした生成AIチャットボットの場合、回答を生成する段階で曖昧な情報が少なく、曖昧な返答をする以前に、「返答しない」（＝できない）パターンが多くなったと考えられた。

ハルシネーションは一般公開されているChatGPTおよびGoogle検索結果を参照情報とした生成AIチャットボットでおおよそ10%程度確認された。一方、がん情報サービスを参照情報として回答を生成した場合は、ハルシネーションを認めなかった。前者は、返答を生成する段階での参照情報にエビデンスに基づかない情報が含まれていた可能性が大きく、特にGoogle検索の結果、宣伝・広告サイト内のテキストデータも含まれている。また、一般公開されているChatGPT-3.5の学習データも詳細は明らかにされていない。生成型人工知能自体は情報の真偽は判断しないため、参照情報に誤情報が含まれていてもそのまま文章を生成し[6]、結果的に回答にハルシネーションが含まれると考えらえれた。逆に正しい情報＝「がん情報サービス」が参照情報の場合はデータベース上には正しい情報のみ存在する。よって生成された回答もハルシネーションを認めることはなかったと考えられた。

確かながんに関する情報群を参照情報として限定した生成AIチャットボットは、ハルシネーションを減少できる利点を認めた。しかし、科学的エビデンスのない情報に対してその説明や理由を返答することはなく、「情報が存在しないため返答できない」と返答し、結果的に返答できる範囲が限られた。一般向けシステムとしては、より広い範囲の質問に対して返答する必要がある。これらの解消には、参照する情報源となるWebサイト自体のテキストの工夫や質問内容に対するプロンプトの改良が必要と考えられた。

E. 結論

確かながんに関する情報群を参照情報として限定

した生成AIチャットボットは、ハルシネーションを減少できる利点を有する事から、長大な情報の中からエビデンスに基づいた科学的に妥当な情報源を得る手段として利用できる可能性がある。一般向けシステムとして、参照する情報源となるWebサイト自体のテキストの工夫や質問内容に対するプロンプトの改良が現状では必要である。

F. 健康危険情報

特になし

G. 研究発表

（発表誌名巻号・頁・発行年等も記入）

1. 論文発表　なし

2. 学会発表　なし

H. 知的財産権の出願・登録状況

（予定を含む）

1. 特許取得　なし

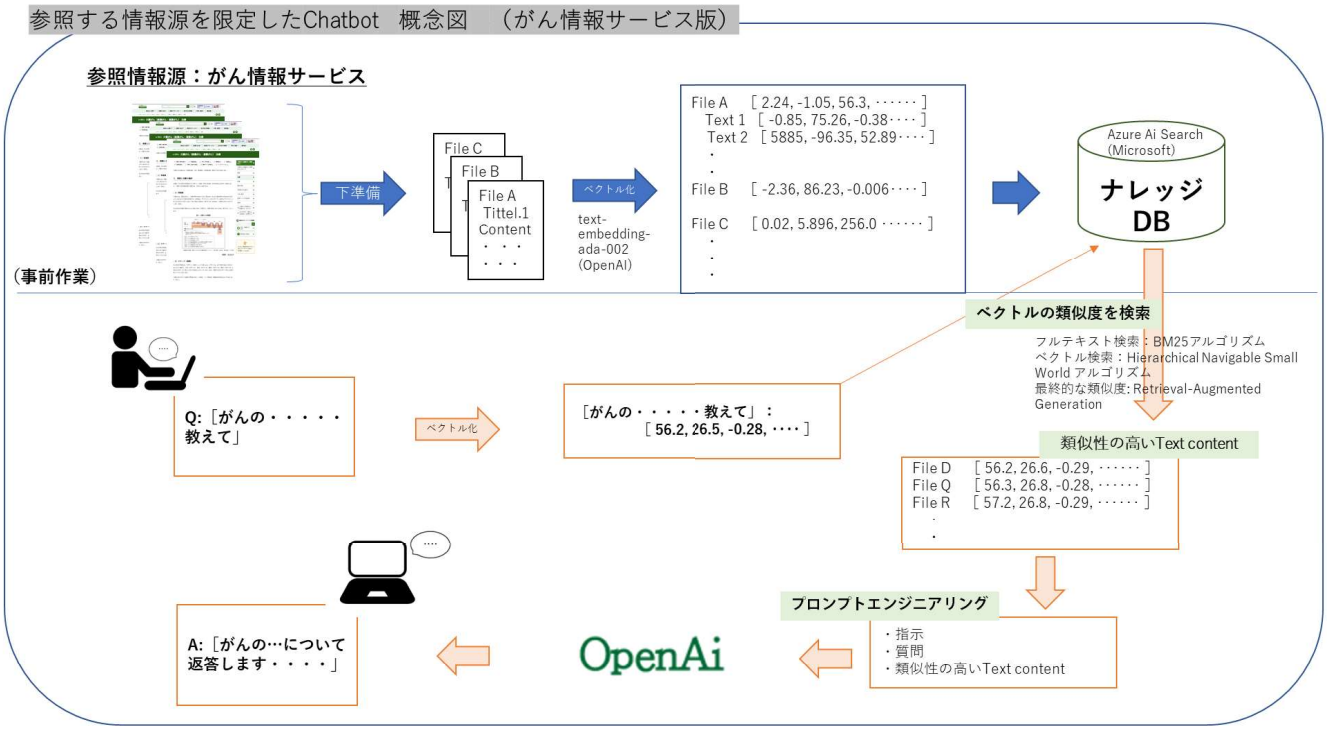
2. 実用新案登録 なし

3. その他　なし

資料

（A）情報源：ChatGPT自体

（B）情報源：がん情報サービス



（C）情報源：Google検索

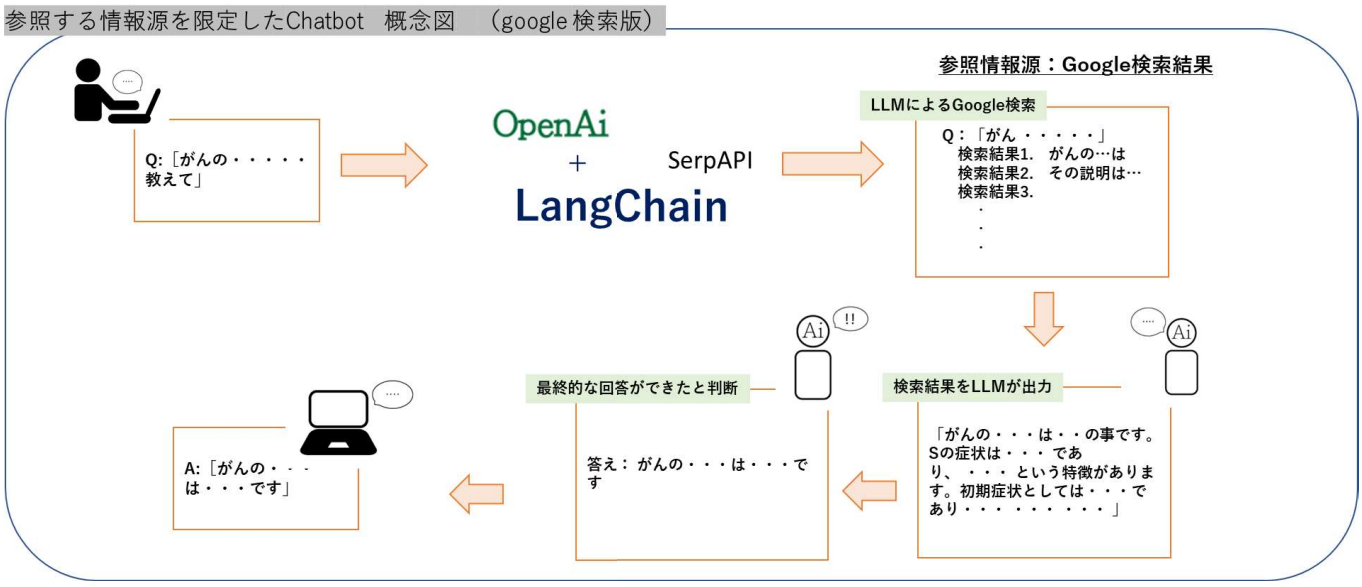


図１．参照する情報を限定した生成AI Chatbot　概念図．（A）情報源：ChatGPT自体．（B）情報源：がん情報サービス．（C）情報源：Google検索

（A）

（B）

図2. 参照する情報を限定した生成AI Chatbot　インターフェイス.　（A）情報源：がん情報サービス.（B）情報源：Google検索

表1. 生成AI Chatbotごとの質問に対する返答の有無

生成AI Chatbot /モデル	全体での集計 (N = 40)				がん情報サービスに存在する情報での質問 (*n* = 20)				がん情報サービスに存在しない情報での質問 (*n* = 20)			
	あり		なし		あり		なし		あり		なし	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Conv. ChatGPT-3.5	40	100	0	0	20	100	0	0	20	100	0	0
がん情報GPT-4	14	35.0	26	65.0	14	70.0	6	30.0	0	0	20	100
がん情報GPT-3.5	19	47.5	21	52.5	19	95.0	1	5.0	0	0	20	100
Google GPT-4	23	57.5	17	42.5	11	55.0	9	45.0	12	60.0	8	40.0
Google GPT-3.5	34	85.0	6	15.0	19	95.0	1	5.0	15	75.0	5	25.0
p value (chi-square test)	$P < 0.001$				$P < 0.001$				$P < 0.001$			

表2. 生成AI Chatbotごとの返答ない場合の追加情報の有無

生成AI Chatbot /モデル	全体での集計 (N = 40)				がん情報サービスに存在する情報での質問 (*n* = 20)				がん情報サービスに存在しない情報での質問 (*n* = 20)			
	あり		なし		あり		なし		あり		なし	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Conv. ChatGPT-3.5	0	0	0	0								
がん情報GPT-4	4	15.4	22	84.6	1	16.7	5	83.3	3	15.0	17	85.0
がん情報GPT-3.5	11	52.4	10	47.6	1	100	0	0	10	50.0	10	50.0
Google GPT-4	0	0	17	100	0	0	9	100	0	0	8	100
Google GPT-3.5	1	16.7	5	83.3	0	0	1	100	1	20.0	4	80.0
p value (chi-square test)	$P = 0.001$				$P = 0.03$				$P = 0.02$			

表3. 生成AI Chatbotごとの質問に対する曖昧な返答の有無

生成AI Chatbot /モデル	全体での集計 (N = 40)				がん情報サービスに存在する情報での質問 (*n* = 20)				がん情報サービスに存在しない情報での質問 (*n* = 20)			
	あり		なし		あり		なし		あり		なし	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Conv. ChatGPT-3.5	4	10.0	36	90.0	1	5.0	19	95.0	3	15.0	17	85.0
がん情報GPT-4	0	0	14	100	0	0	14	100	0	0	0	0
がん情報GPT-3.5	0	0	19	100	0	0	19	100	0	0	0	0
Google GPT-4	5	21.7	18	78.3	2	18.2	9	81.8	3	25.0	9	75.0
Google GPT-3.5	9	25.7	26	74.3	3	15.8	16	84.2	6	37.5	10	62.5
p value (chi-square test)	$P = 0.02$				$P = 0.15$				$P = 0.30$			

表4. 生成AI Chatbotごとの質問に対する対するハルシネーションの有無

生成AI Chatbot /モデル	全体での集計 (N = 40)				がん情報サービスに存在する情報での質問 (*n* = 20)				がん情報サービスに存在しない情報での質問 (*n* = 20)			
	あり		なし		あり		なし		あり		なし	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Conv. ChatGPT-3.5	3	7.5	37	92.5	1	5.0	19	95.0	2	10.00	18	90.0
がん情報GPT-4	0	0	14	100	0	0	14	100	0	0	0	0
がん情報GPT-3.5	0	0	19	100	0	0	19	100	0	0	0	0
Google GPT-4	3	13.0	20	87.0	0	0	11	100	3	25.0	9	75.0
Google GPT-3.5	4	11.8	30	88.2	0	0	19	100	4	26.7	11	73.3
p value (chi-square test)	$P = 0.35$				$P = 0.53$				$P = 0.39$			

引用文献

1. Chen X, Siu LL. Impact of the Media and the Internet on Oncology: Survey of Cancer Patients and Oncologists in Canada. *JCO.* 2001;19(23):4291-4297.
2. Ogasawara R, Katsumata N, Toyooka T, Akaishi Y, Yokoyama T, Kadokura G. Reliability of Cancer Treatment Information on the Internet: Observational Study. *JMIR cancer.* 2018;4(2):e10031.
3. Hill S, Mao J, Ungar L, Hennessy S, Leonard CE, Holmes J. Natural supplements for H1N1 influenza: retrospective observational infodemiology study of information and search activity on the Internet. *J Med Internet Res.* 2011;13(2):e36.
4. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nature Medicine.* 2022;28(1):31-38.
5. Bernstein IA, Zhang YV, Govil D, et al. Comparison of Ophthalmologist and Large Language Model Chatbot Responses to Online Patient Eye Care Questions. *JAMA Netw Open.* 2023;6(8):e2330320.
6. Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science. *Nature.* 2023;614(7947):214-216.