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）

Chatbotによるがんの情報提供　-汎用性の高いシステムをめざして-

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）
研究協力者　Mingxin Liu　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、国立がん研究センター「がん情報サービス」を参照情報源とする Retrieval-Augmented Generation（RAG）型AIチャットボットを構築し、複数の大規模言語モデル（LLM；GPT-4、Claude-3.5、Gemini-2.0）および異なる知識利用条件における応答特性を比較検討した。特に、「参照情報のみに基づく応答（RAG-only）」と、「参照情報に加えてLLM内部知識を併用した応答（Hybrid RAG）」について、応答可能性およびhallucination発生との関連を解析した。評価には、がん情報サービス内に情報が存在する質問31問、および存在しない質問31問の計62問を使用した。生成された372件の回答について、Grounded Response、Unsupported Response、Unclear Response、Containing Hallucinations、No Answerの5分類による評価を行った。その結果、RAG-only条件ではhallucination発生率は低く（3–6%）、高い安全性が認められた。一方、回答不能（No Answer）が増加し、特に参照情報に存在しない質問では有意に回答生成率が低下した。これに対し、Hybrid RAG条件では応答生成率が向上し、Claude-3.5およびGPT-4では98%の質問に回答を生成した。しかし、hallucination発生率は16–24%へ増加し、参照情報に存在しない内容についてLLM内部知識を用いて補完的回答を生成する過程で誤情報が発生していた。多変量解析では、hallucination発生にはLLM種類よりも知識利用条件および回答長が強く関連していた。一方、回答不能は、参照情報に存在しない質問およびRAG-only条件で有意に増加した。以上より、RAG型AIチャットボットによるがん情報提供では、「hallucination抑制」と「応答可能性」の間に明確なトレードオフが存在することが示された。今後、安全性と実用性を両立した医療AIシステムを実現するためには、LLM性能のみならず、知識利用方法、回答制御、出力長を含めた統合的なシステム設計が重要であると考えられる。

A．研究目的

近年、インターネットおよびソーシャルメディアの発達により、患者は医療情報へ容易にアクセスできるようになり、治療選択や意思決定に活用する機会が増加している。一方で、Web上には不正確あるいは誤解を招く情報も多数存在しており、特にがん領域では、誤った情報に基づく判断が適切な治療機会の喪失や経済的負担の増大につながることが問題とされている。

近年では、大規模言語モデル（Large Language Model：LLM）を用いた生成AIチャットボットが医療情報提供手段として注目されている。AIチャットボットは、従来の検索エンジンと比較して自然言語による対話的な情報取得が可能であり、ユーザーにとって理解しやすく利便性が高いとされている。しかし、生成AIには事実に基づかない情報を生成する hallucination の問題が存在し、医療分野では安全性の観点から大きな課題となっている。

これに対し、外部情報を参照しながら回答を生成する Retrieval-Augmented Generation（RAG）は、信頼性の高い情報源に基づく回答生成を可能とし、hallucination抑制手法として注目されている。これまで我々は、がん情報サービス（Cancer Information Service：https://ganjoho.jp）を参照情報源としたRAG型AIチャットボットにより、hallucinationの生成割合を大幅に低減できることを報告してきた[1]。一方で、参照情報のみに基づいて回答を生成する場合、情報源に含まれない質問に対しては回答不能となりやすく、

応答可能性や汎用性が低下するという課題も明らかとなった。すなわち、RAG型AIチャットボットにおいては、hallucination抑制と応答可能性の間にトレードオフが存在すると考えられる。しかし、どのような設計条件がこのバランスに影響を与えるのかについては、十分な検討が行われていない。特に、異なるLLM間での応答特性の差異や、参照知識の範囲の違いが、hallucinationおよび応答不能に与える影響は明らかではない。

そこで本研究では、がん情報提供を目的としたRAG型AIチャットボットを構築し、複数のLLMおよび異なる知識利用条件を用いて応答特性を比較した。これにより、hallucinationを抑制しつつ、応答可能性および情報網羅性を両立するための設計条件を明らかにし、安全かつ実用的ながん情報提供AIシステム構築に資することを目的とした。

B. 研究方法

RAG型AIチャットボットの構築

正確な情報を参照しながら回答を生成するRAG型AIチャットボットを構築し、生成条件の違いによる応答特性を比較検討した（図1）。参照情報源には、国立がん研究センターが運営する「がん情報サービス」のWebサイト全体のテキストデータを用いた。Web上のコンテンツを抽出・整形し、ベクトル化した後、ナレッジデータベースとして格納した。ユーザーからの質問に対しては、ベクトル類似度検索により意味的関連性の高いテキストを抽出し、その内容を参照情報として利用した。得られた参照情報は、プロンプトエンジニアリングを用いて大規模言語モデル (LLM) へ入力し、回答を生成させた。システム全体はPythonを用いて構築し、Google Colaboratory環境上で実装した。本システムの環境構築および開発は、生成AI開発企業（株式会社pipon、東京都）へ委託した。

生成条件の設定

生成条件として、（1）LLMの違い、および（2）参照情報利用条件の違いを検討した（図1）。LLMとして、GPT-4（OpenAI）、Gemini-2.0-flash-001（Google DeepMind）、Claude-3.5-sonnet-20240620（Anthropic）の3種類を使用した。また、参照情報利用条件として、以下の2条件を設定した。

1．RAG-only条件：取得した参照情報のみに基づいて回答を生成する条件。2．Hybrid RAG条件：取得した参照情報に加え、LLM内部知識（parametric knowledge）も利用して回答を生成する条件。以上より、6種類の生成条件（RAG-only Claude、RAG-only GPT、RAG-only Gemini、Hybrid RAG Claude、Hybrid RAG GPT、Hybrid RAG Gemini）を比較対象とした。

質問文の作成

質問文は、一般利用者が実際に抱えるがん情報ニーズを反映することを目的として作成した。質問候補は、「がん情報サービス」およびYahoo!知恵袋から収集し、日常的な疑問や治療・生活に関する内容を中心に整理した。収集した質問について、内容の重複除外および文章構造の統一化を行い、最終的に日本語による62問を評価用データセットとして使用した。このうち31問はがん情報サービス内に対応情報が存在する質問、残る31問はがん情報サービスに対応情報が存在しない質問とした。各質問文には、「参照情報に基づいて回答し、情報が存在しない場合にはその旨を示すこと」とする統一プロンプトを付与した上で、各条件下でAIチャットボットへ入力した。

応答内容の評価

生成された回答について、研究者1名が内容評価を行った。判定困難例については、がん情報提供経験を有する医師と協議の上で判定した。回答は以下の5分類に整理した；1．Grounded Response; 参照情報に基づいた適切かつ正確な回答, 2．Unsupported Response; 質問に直接回答せず、一般的助言や受診勧奨を行う回答, 3．Unclear Response; 曖昧あるいは非断定的表現を含む回答, 4．Containing Hallucinations; 誤情報または根拠不明情報を含み、医療判断へ影響を与える可能性がある回答, 5．No Answer; 情報不足等を理由として回答生成を行わなかったもの

統計解析

各条件について、回答生成率、回答文字数、Grounded Responseの割合、hallucination割合を算出した。さらに、「No Answer」に関連する因子、および「Containing Hallucinations」に関連する因子を解析した。候補因子として、質問内容（がん情報サービス内/外）、LLM種類、知識利用条件（RAG-only/Hybrid RAG）、出力文字数（Token）を用いた。単変量解析には Fisher’s exact test を用い、その後、多変量ロジスティック回帰分析を行い、調整オッズ比（Odds Ratio；OR）および95%信頼区間（95% C

I）を算出した。統計解析には Stata/IC version 19.0（StataCorp LP, College Station, TX, USA）を使用し、有意水準は両側P < 0.05とした。

（倫理面への配慮）

本研究は公開情報のみを対象としており、個人情報は取り扱わないため、倫理上の問題は生じないと考えられる。

C. 研究結果

1. チャットボットの応答パターン

全62問に対し、6条件（3種類のLLM × 2種類の知識利用条件）で回答生成を行い、合計372件の応答を解析した（図2）。372件中、273件（73%）で回答生成が行われ、99件（27%）では回答生成が行われなかった。

生成された回答のうち、179件（全体比；48%）はGrounded Responseに分類され、参照情報に基づき質問へ直接回答しており、医学的内容についても正確かつ具体的な説明が認められた（表1）。逆に、48件（13%）はContaining Hallucinationsに分類された。参照情報に存在しない内容について、LLM内部知識を用いて補完的に回答を生成する傾向が認められた。Unsupported Responseは32件（9%）であり、質問に対する直接的回答は行わず、「医療機関へ相談することを推奨する」「がん相談支援センターへ相談する」といった一般的助言を中心とした応答が認められた。Unclear Response 14件（4%）であり、「明確ではない」「可能性がある」等の曖昧あるいは非断定的表現が多く認められた。No Answerは、99件（27%）であった。

2. チャットボット条件ごとの応答性能

各条件における応答生成率、回答長、回答品質を比較した（表2）。RAG-only条件では、Claude-3.5が最も高い応答生成率を示し、62問中51問（82%）に回答した。一方、GPT-4では22問（35%）、Gemini-2.0では30問（48%）にとどまった。Hybrid RAG条件では、すべてのLLMで応答生成率が上昇し、Claude-3.5およびGPT-4では61問（98%）、Gemini-2.0では48問（77%）に回答した。Grounded Response割合についても条件間で差を認め、Hybrid RAG Claude が最も高く、62問中45問（73%）で適切な回答を生成した。続いてHybrid RAG GPT（53%）、Hybrid RAG Gemini（50%）であった。一方、Containing Hallucinations の割合はHybrid RAG条件で増加し、Hybrid RAG Claudeでは14件（23%）、Hybrid RAG GPTでは15件（24%）、Hybrid RAG Geminiでは10件（16%）に認められた。RAG-only条件ではhallucination割合は低く、RAG-only Claude 6%、RAG-only GPT 3%、RAG-only Gemini 5%であった。また、Hybrid RAG条件では回答文が長文化する傾向を認め、特にHybrid RAG Claudeでは中央値666 tokensと最も長い出力を示した。

3. 「No Answer」に関連する因子

回答不能（No Answer）に関連する因子を解析した（表3）。多変量ロジスティック回帰分析の結果、「がん情報サービス」に情報が存在しない質問は、存在する質問と比較して有意に回答不能となりやすかった（OR 16.4、95% CI 7.6–35.3、$P < 0.001$）。LLM別では、Claude-3.5と比較して、GPT-4（OR 11.3）およびGemini-2.0（OR 15.4）で回答不能割合が有意に高かった。さらに、Hybrid RAG条件と比較し、RAG-only条件では回答不能となる割合が有意に高く（OR 22.5、95% CI 10.2–49.7、$P < 0.001$）、参照情報のみを利用する構造では応答可能性が低下することが示された。

4. hallucination生成に関連する因子

hallucination生成に関連する因子を解析した（表4）。多変量解析の結果、LLM種類自体はhallucination発生に有意な関連を認めなかった。一方、知識利用条件は有意に関連しており、Hybrid RAG条件と比較して、RAG-only条件ではhallucination発生リスクが低下していた（OR 0.4、95% CI 0.2–0.9、$P = 0.03$）。また、長文回答（375–1178 tokens）は、短文回答（29–374 tokens）と比較してhallucination発生リスクが有意に高かった（OR 3.0、95% CI 1.2–7.3、$P = 0.01$）。

D. 考察

本研究では、国立がん研究センター「がん情報サービス」を参照情報源とするRAG型AIチャットボットを構築し、複数の大規模言語モデル（LLM）および異なる知識利用条件における応答特性を比較検討した。その結果、RAG-only条件ではhallucinationを低減できる一方、回答不能（No Answer）が増加し、Hybrid RAG条件では応答可能性が向上する一方でhallucinationリスクが増加するという、明確なトレードオフが存在することが示された。

本研究において、RAG-only条件ではhallucination発生率は低く、参照情報を信頼性の高い「がん情報サービス」に限定することで、生成される回答の安全性を高められることが示唆された。これは、外部情報を参照しながら回答を生成するRAG構造が、医療分野

におけるhallucination抑制に有効であることを支持する結果と考えられる。一方で、RAG-only条件では、参照情報に十分な記載が存在しない場合に回答不能となるケースが増加し、特に「がん情報サービス」に情報が存在しない質問では有意にNo Answerが増加した。このことは、外部知識のみに依存する構造では、知識ベースの網羅性が応答可能性を大きく左右することを示している。一方、Hybrid RAG条件では、LLM内部知識を併用することで応答生成率が大幅に向上した。しかしその反面、hallucination発生率も増加した。特に、参照情報に存在しない内容について、LLMが内部知識を用いて補完的に回答生成を行う過程でhallucinationが発生していた。これらの結果は、「応答可能性」と「安全性」の間に構造的トレードオフが存在することを示している。

本研究では、LLM種類そのものはhallucination発生と有意な関連を認めなかった。一方で、知識利用条件（RAG-only/Hybrid RAG）や回答長はhallucinationと有意に関連していた。この結果は、医療AIシステムの安全性は、単純なLLM性能差よりも、「どのように知識を利用するか」というシステム設計に強く依存する可能性を示唆している。特に、長文回答ではhallucinationリスクが増加しており、情報量の増加とともに、LLMが不確実な内容を補完的に生成しやすくなる可能性が考えられた。

また、本研究ではNo Answerの挙動について重要な知見が得られた。従来、回答不能はAI性能の低下として扱われることが多いが、医療分野では、不確実な状況下で無理に回答を生成しないこと自体が安全性に寄与する可能性がある。実際、本研究では、Claude-3.5は高い応答生成率を示した一方でhallucinationも増加し、GPT-4やGemini-2.0では、情報不足時に回答生成を控える傾向が認められた。これらの違いは、各LLMの知識量そのものよりも、「不確実性への応答戦略」の違いを反映している可能性がある。さらに、本研究の結果から、医療AIシステムにおいては、「常に回答すること」が最適とは限らないことが示唆された。特にがん領域では、患者の心理的負担や医療安全への影響を考慮すると、不確実な情報を断定的に提示することは重大な問題となりうる。そのため、今後の医療AI設計においては、hallucination抑制、応答可能性、回答長、透明性を統合的に制御する設計思想が重要になると考えられる。

本研究は、がん情報提供AIにおける「安全性」と「汎用性」のバランスを定量的に評価した基礎的検討であり、今後の医療AIチャットボット開発における重要な設計指針を示したものと考えられる。

E. 結論

本研究により、RAG型AIチャットボットによるがん情報提供において、「hallucination抑制」と「応答可能性」の間には明確なトレードオフが存在することが示された。RAG-only条件ではhallucinationを低減できる一方、回答不能が増加し、Hybrid RAG条件では応答可能性が向上する一方でhallucinationリスクが増加した。また、hallucination発生にはLLM種類そのものよりも、知識利用方法や回答長などのシステム設計要因が強く関連していることが示唆された。特に、参照情報に基づく回答制御および適切な応答長管理は、安全性向上に重要であると考えられた。以上より、今後の医療AIチャットボット開発においては、単純なLLM性能向上だけではなく、知識利用方法、回答制御、不確実性への対応方針を含めた統合的なシステム設計が重要であると考えられる。これらを適切に最適化することで、安全性と実用性を両立した、汎用性の高いがん情報提供AIシステムの実現が期待される。

F. 健康危険情報

特になし

G. 研究発表

1. 書籍発表　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

参考文献

1. Nishisako S, Higashi T, Wakao F. Reducing Hallucinations and Trade-Offs in Responses in Generative AI Chatbots for Cancer Information: Development and Evaluation Study. *JMIR Cancer*. Sep 11 2025;11:e70176. doi:10.2196/70176

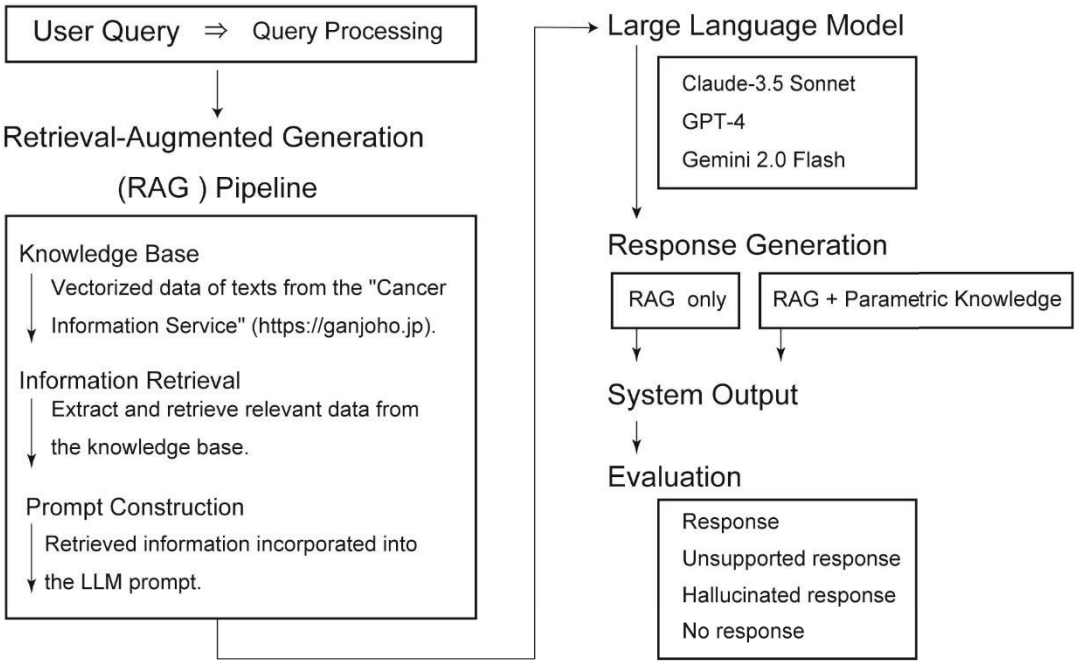
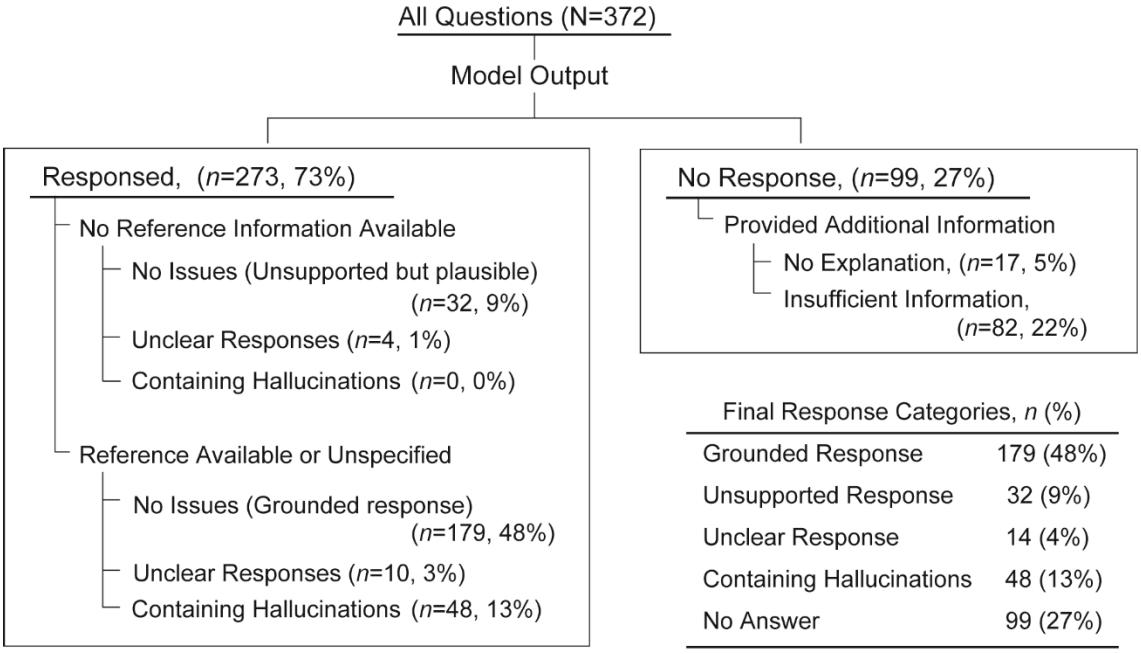


図1. 本研究の全体像



Final Response Categories, *n* (%)	
Grounded Response	179 (48%)
Unsupported Response	32 (9%)
Unclear Response	14 (4%)
Containing Hallucinations	48 (13%)
No Answer	99 (27%)

図2. チャットボットの応答パターン

表1．異なるモデル構成におけるチャットボット応答の代表例およびその分類[a]

質問	構成	応答例	分類
胃がんは咳やくしゃみでうつると聞きました。本当ですか？	RAG-only (Claude) [b]	“提供された情報によると、胃がんは感染症ではなく、咳やくしゃみによって他人へうつることはありません。胃がんは、胃粘膜細胞の異常増殖によって発生する疾患であり、Helicobacter pylori（ピロリ菌）感染、喫煙、高塩分摂取などがリスク因子として知られています。”	Grounded response
がんに最も効果のあるキノコを教えてください。	Hybrid RAG (Claude) [c]	“がんに最も効果があると断定できるキノコはありません。一部のキノコには免疫機能を補助する可能性が示唆されていますが、標準的ながん治療の代替となるものではありません。”	Grounded response
「札幌市北区に住んでいます。知人が胆管がんと診断されました。治療を受ける場合、どの病院が勧められますか？」	RAG-only (GPT)	“提供された文章には、札幌市北区における推奨病院に関する情報は含まれていません。”	No response (insufficient information)
	RAG-only (Gemini)	“I do not know.”	No response (no explanation)
	RAG-only (Claude)	“特定の病院を推奨することはできませんが、がん相談支援センターや大学病院等へ相談することが考えられます。”	Unsupported response
	Hybrid RAG (GPT)	“提供された文章には具体的な病院情報は含まれていませんが、大学病院などの地域医療機関へ相談することが考えられます。”	Unsupported response
	Hybrid RAG (Gemini)	“提供された文章には病院情報は含まれていませんが、胆管がんの治療には手術や化学療法などが含まれる場合があります。”	Unsupported response
	Hybrid RAG (Claude)	“北海道大学病院には、がん治療センターが設置されています。また、手稲渓仁会病院にも、がん治療センターが設置されています。”	Containing Hallucinations

[a]代表的な応答例は評価用データセットから抽出し、Grounded Response; 参照情報に基づいた適切かつ正確な回答, Unsupported Response; 質問に直接回答せず、一般的助言や受診勧奨を行う回答, Unclear Response; 曖昧あるいは非断定的表現を含む回答, Containing Hallucinations; 誤情報または根拠不明情報を含み、医療判断へ影響を与える可能性がある回答, No Answer; 情報不足等を理由として回答生成を行わなかったものに分類した。

[b]本研究において、ClaudeはClaude 3.5、GPTはGPT-4、GeminiはGemini 2.0を指す。

[c] Hybrid RAGは、検索拡張生成（retrieval-augmented generation）と大規模言語モデルのパラメトリック知識を組み合わせた構成を示す。

表2. チャットボット条件ごとの応答性能

構成 [a]	応答生成数 N = 273, *n* (%)	生成文字数 (token, range)	Grounded response[b] N=179, *n* (%)	Unsupported Response N = 32, *n* (%)	Unclear Response N=14, *n* (%)	Containing Hallucinations N=48, n (%)
RAG-only (Claude) *n* = 62	51 (82)	424 (155-794)	27 (44)	14 (22)	6(10)	4 (6)
RAG-only (GPT) *n* = 62	22 (35)	163.5 (41-662)	18 (29)	1(2)	1(2)	2 (3)
RAG-only (Gemini) *n* = 62	30 (48)	138 (30-648)	25 (40)	1(2)	1(2)	3 (5)
Hybrid RAG (Claude) *n* = 62	61 (98)	666 (477-1178)	45 (73)	0 (0)	2 (3)	14 (23)
Hybrid RAG (GPT) *n* = 62	61 (98)	247 (106-965)	33 (53)	12 (21)	1 (2)	15 (24)
Hybrid RAG (Gemini) *n* = 62	48 (77)	224 (29-1178)	31 (50)	4 (6)	3 (5)	10 (16)
***p* value[c]**	< 0.001	< 0.001	< 0.001	0.82	0.18	< 0.001

a Hybrid RAGは、検索拡張生成（retrieval-augmented generation）と大規模言語モデルのパラメトリック知識を組み合わせた構成を示す。本研究において、ClaudeはClaude 3.5、GPTはGPT-4、GeminiはGemini 2.0を指す。

b Grounded Response; 参照情報に基づいた適切かつ正確な回答, Unsupported Response; 質問に直接回答せず、一般的助言や受診勧奨を行う回答, Unclear Response; 曖昧あるいは非断定的表現を含む回答, Containing Hallucinations; 誤情報または根拠不明情報を含み、医療判断へ影響を与える可能性がある回答, No Answer; 情報不足等を理由として回答生成を行わなかったもの

c Fisherの正確確率検定を用いて解析した。

表3. 回答不能（No Answer）に関連する因子（N=372）

因子	単変量解析 [a] *n* (%)	*p* 値	多変量解析[b] OR[c] (95% CI)	*p* 値
応答が生成されなかった質問	99 (27)			
質問内容		< 0.001		
CISに掲載あり[d]	20 (11)		1 (reference)	
CISに掲載なし	79 (42)		16.4 (7.6-35.3)	< 0.001
LLM種類[e]		< 0.001		
Claude	12 (10)		1 (reference)	
GPT	41 (33)		11.3 (4.5-28.4)	< 0.001
Gemini	46 (37)		15.4 (6.0-39.5)	< 0.001
知識利用状況[f]		< 0.001		
Hybrid RAG	16 (9)		1 (reference)	
RAG-only	83 (45)		22.5 (10.2-49.7)	< 0.001

[a] Fisherの正確確率検定を用いて解析した

[b]多変量ロジスティック回帰分析を用いて解析した

[c] OR: odds ratio

[d] CIS: Cancer Information service;「がん情報サービス」

[e] Claude refers to Claude 3.5, GPT refers to GPT-4, and Gemini refers to Gemini 2.0.

[f] Hybrid RAGは、検索拡張生成（retrieval-augmented generation）と大規模言語モデルのパラメトリック知識を組み合わせた構成を示す。本研究においては、Hybrid RAG構成を知識ソースとして用いた

表4. hallucination生成に関連する因子

因子	単変量解析 [a] *n* (%)	*p* 値	多変量解析[b] OR[c] (95% CI)	*p* 値
Hallucinationを含む応答が生成された質問	48 (13)			
質問内容		0.09		
CISに掲載あり[d]	30 (16)		1 (reference)	
CISに掲載なし	18 (10)		0.99 (0.5-2.0)	0.97
LLM種類[e]		0.62		
Claude	18 (15)		1 (reference)	
GPT	17 (14)		2.4 (0.9-6.2)	0.08
Gemini	13 (11)		1.9 (0.7-5.0)	0.20
知識利用状況[f]		< 0.001		
Hybrid RAG	39 (21)		1 (reference)	
RAG-only	9 (5)		0.4 (0.2-0.9)	0.03
出力文の長さ (tokens)		0.04		
29-374	17 (13)		1 (reference)	
375-1178	31 (23)		3.0 (1.2-7.3)	0.01

[a] Fisherの正確確率検定を用いて解析した

[b]多変量ロジスティック回帰分析を用いて解析した

[c] OR: odds ratio

[d] CIS: Cancer Information service;「がん情報サービス」

[e] Claude refers to Claude 3.5, GPT refers to GPT-4, and Gemini refers to Gemini 2.0.

[f] Hybrid RAGは、検索拡張生成（retrieval-augmented generation）と大規模言語モデルのパラメトリック知識を組み合わせた構成を示す。本研究においては、Hybrid RAG構成を知識ソースとして用いた