厚生労働科学研究費補助金（がん対策推進総合研究事業）
科学的根拠に基づくがん情報の提供及び均てん化に向けた体制整備に資する研究（23EA1026）
（分担研究報告書）

正しい情報源を参照し返答する生成 AI によるがんの情報提供
～ハルシネーションと返答割合のトレードオフ～

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

本研究ではがんに関する正確な医療情報を提供する生成 AI chatbot を開発する目的において、正しいがんの情報 (＝がん情報サービス, CIS) を AI に参照させた際の返答の特徴を把握する。検索拡張生成 (RAG: Retrieval-Augmented Generation) を用いて、限定的に情報を参照し返答する Chatbot を試作した。参照情報源は、がん情報サービス (CIS Chatbot) および Google 検索結果 (Google Chatbot) のテキストデータであり、大規模言語モデルは GPT-4 および -3.5 (-turbo-16k) とした。CIS に含まれる内容および含まれないがん関連の質問 (日本語) に対する返答を従来型 Chatbot (参照先を指定しない) 含め分析した。ハルシネーション生成の割合は、従来型 Chatbot ではおおよそ 40%であったのに対し、CIS Chatbot の場合 GPT-4 で 0%・GPT-3.5 で 3%、Google Chatbot ではそれぞれ 13%・23%であった。参照情報源をハルシネーション生成の因子として比較した際、参照先を指定しない場合、CIS の参照に対しその可能性は高くそのオッズは 16.1 (95% CI, 3.7-50.0) であった。従来型 Chatbot はすべての質問に返答したが、RAG 機能を持つ Chatbot ではその返答率は減少し (36-81%)、さらに CIS Chatbot は CIS に含まれない内容の質問には返答しなかった (0%)。確かながんに関する情報群を参照情報として限定した Chatbot は、ハルシネーションを大幅に減少させる事ができるが、同時に科学的 Evidence のない情報に対して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。Evidence に基づきにくい事柄をどのように回答させるかを考えこのトレードオフを解消する事により、誤情報から患者を守るべく RAG-生成 AI による正確な医療情報の提供が可能となる。

A. 研究目的

インターネット上に広がる誤情報は人々の健康に深刻な悪影響を与える可能性があり、特に医療分野においては正確な情報提供が強く求められている。現在では多くの患者が健康に関する情報をインターネットから取得しており、誤った情報に基づいた判断を下すリスクを含んでいる。こうした課題に対し、人工知能（AI）の技術が医療への応用において注目されている。すでに放射線科、病理学、消化器科、眼科などでの画像診断の支援に活用されており、近年では自然言語処理を用いた生成型のAI Chatbot（生成AI）が、患者への情報提供手段としても利用されつつある。生成AIは、従来の検索エンジンと比べて対話的かつわかりやすい形で情報を提供できる点が利点とされている。一方で、生成AIには「ハルシネーション」と呼ばれる問題が存在する。これは、もっともらしいが事実と異なる情報を出力してしまう現象であり、医療分野のように高い正確性が求められる領域では大きな課題である。

ハルシネーションの軽減を目的として「検索拡張生成（RAG：Retrieval-Augmented Generation）」という手法が開発されている。RAGは、AIの事前学習データだけでなく、外部の非パラメトリックメモリ（専門データベースやウェブ上の情報など）を活用し、最新かつ関連性の高い情報に基づいた応答を生成することで、ハルシネーションの発生を抑えることが可能である。我々は確かながん情報を提供する代表的な医療情報サイト「がん情報サービス（CIS, https://ganjoho.jp/）」を情報源としたRAG-AI Chatbotは正確な医療情報を提供できるのではないかと仮定した。本研究では、将来的な生成AIによる正確な医療情報提供を目指し、確かな情報を参照データ

としたChatbotの返答の特徴を把握することを目的とした。なお本報告は、令和5年度 科学的根拠に基づくがん情報の提供及び均てん化に向けた体制整備に資する研究 (23EA1026 : 研究代表者 若尾 文彦) 報告中「参照する情報源を限定した生成AI Chatbotによるがん情報提供のハルシネーション排除の可能性」に対してデータを追加した上で再解析を実施したものを記載する。

B. 研究方法

限定的に情報を参照し返答するRAG-AI Chatbotを試作した。CISに含まれる内容および含まれないがん関連の質問 (日本語) に対する返答を従来型Chatbot (参照先を指定しない) 含め分析した。比較対象としたのは、①国立がん研究センターが運営する「がん情報サービス」を情報源とするRAGシステムを有するChatbot (CIS Chatbot)、②Google検索結果のテキストデータを参照するChatbot (Google Chatbot)、③学習済みデータのみで応答する従来型Chatbot（Conventional Chatbot）の3種類である。大規模言語モデル (LLM) はGPT-4 (Generative Pre-trained Transformer, OpenAI) および -3.5 (-turbo-16k) を用いた。CISに含まれる内容（31問）および含まれない質問の合計62件の質問を6種のChatbot（3種×GPT-4/GPT-3.5）に入力し、計372件の応答を収集した。

Chatbotからの返答は「返答なし」「補足情報あり返答なし」「あいまいな返答」「ハルシネーションを含む応答」「問題なし」の5つに分類した。返答有無（回答の割合）及びハルシネーションを含む返答の割合を質問の性質（CISに含まれる内容/含まれない内容）別に集計した。ハルシネーションを含む返答に寄与する因子の解析として一般化線形混合効果モデルを用いて因子解析を実施した。

（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. Chatbotの応答傾向と返答例

合計372件の質問への応答は、31%は「返答なし」、5%は「補足情報あり返答なし」と分類された。返答の有った全体の69%のうち、10%が「あいまいな返答」、19%が「ハルシネーションを含む応答」であった。全体の40%は問題のない応答であった（図2）。それぞれの返答例を図３に示した。

2. AI-Chatbotの質問に対する応答の有無（表1）

回答の割合は6種類のChatbot間で36%から100%までであり、それらには有意に差を認めた。RAGを搭載したChatbot（CIS Chatbot, Google Chatbot）は、従来型モデルよりも回答の割合が低かった。CISに情報が掲載されている質問に対して、CIS Chatbotの応答率はGPT-4で71%、GPT-3.5で97%であったが、CISに記載されていない質問には「情報がないため返答しない」（回答の割合0%）と返した。一方、Google-ChatbotはCISに掲載の有無にかかわらず応答しており、GPT-4では45%と52%、GPT-3.5では90%と71%の回答の割合であった。従来型Chatbotは、LLMのバージョンに関係なくすべての質問に応答した。

3. AI-Chatbotの質問に対するハルシネーションを含む返答の割合（表2）

RAGを搭載したAI Chatbotは、従来型のChatbotよりもハルシネーション回答の割合が低かった。特にCISを参照するChatbotでは、最も低いハルシネーションの生成割合を示した。CISに情報が掲載されている質問に対しては、ハルシネーションの発生は少なく、CIS ChatbotではGPT-4が0%、GPT-3.5が6%、Google Chatbotではそれぞれ6%と10%だった。一方、Google Chatbotの返答において、CISに情報がない質問では、ハルシネーションが増加し、GPT-4で19%、GPT-3.5では35%となった。従来型Chatbotは、全体の約40%の応答にハルシネーションが含まれていた。6種類のChatbot間で、ハルシネーション発生の割合には統計学的に有意な差が認められた（$P < .001$）

4. ハルシネーションを含む応答の生成に関連する要因（表3）

二変量解析の結果、RAGに使用された参照データの違いによって、ハルシネーションの生成割合には有意な差が見られた。最も発生率が低かったのはCIS Chatbotであり、次いでGoogle Chatbot、従来型Chatbotの順であった（それぞれ2%、18%、39%、$P = .03$）。質問内容がCIS内に存在するかしないか・出力テキストの文字数・LLMのバージョンはいずれもハルシネーション生成とは関連しなかった（すべてP

≧ 0.05）。多変量解析では、Google検索結果を参照した場合、CISを参照した場合と比べて、ハルシネーションを発生させる可能性が高く、オッズ比は9.4（95%信頼区間：1.2–17.5、$P < .01$）であった。同様に、従来型のGPTは、CISを参照した場合と比べてオッズ比は16.1（95%信頼区間： 3.7–50.0、$P < .001$）と、さらに高いリスクを示した。LLMのバージョンの違いは、ハルシネーション発生の要因にはならなかった。

D. 考察

本研究では、異なる参照データを用いて生成型AI ChatbotにRAGを適用した結果、信頼できるがんに関する情報を参照情報として組み込むことで、従来のGPTモデルやGoogle検索を参照元とした場合に比べて、ハルシネーションの発生が有意に減少することが示された。一方で、参照に関連する情報がない場合、システムは正しく「参照情報が無いため返答しない」と返すことが確認された。正確な参照データをRAG-AI chatbotに使用することで、AIはエビデンスに基づいた有効な情報を提供できることが分かった。しかし、RAGを用いたChatbotは、参照データに含まれる情報の範囲に応じて、回答可能な質問の幅が制限されることも分かった。一般利用に向けた開発過程では、返答の汎用性を持たせる方法を考案する必要がある。

本研究では、誤情報を避けることに焦点を当てハルシネーションの発生を減らすことを目指した。RAGを用いた非パラメトリックメモリLLM（CIS Chatbot／Google Chatbot）は、従来型のパラメトリックメモリLLM（従来型Chatbot）よりもハルシネーションが少ないことが確認された（2-18% *vs.* 39%、$P = .03$）。Google Chatbotは、CIS Chatbotと比較して9.4倍のオッズでハルシネーションが発生する確率が高いことが示され、非パラメトリック知識源の質の正確性が重要であることが分かった。この結果は、Google Chatbotが証拠に欠けた情報（CISに含まれていない情報）に対して高いハルシネーションの生成割合を示したことからも支持されている。LLMは参照情報の正確性を評価せずに応答を生成するため、RAG情報源の正確性が、医療分野における情報提供において重要であることが分かった。

がん患者からの質問は非常に多岐にわたり、それらは証拠に基づく内容に限られないことが一般的である。本研究では、ハルシネーション生成と応答率の間にトレードオフが見られた。従来型Chatbotはすべての質問に回答したが、RAGを使用したシステムでは応答しない割合が増加した（100% *vs.* 36-81%、$P < .001$）。返答しないことは、少なくとも誤情報を提供しないことであるが、Evidenceに基づきにくい事柄をどのように回答させるかを検討してシステムに反映する必要がある。

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Chatbotは、ハルシネーションを大幅に減少させる事ができるが、同時に科学的Evidenceのない情報に対して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。Evidenceに基づきにくい事柄をどのように回答させるかを考えこのトレードオフを解消することにより、誤情報から患者を守るべくRAG-生成AIによる正確な医療情報の提供が可能となる。

F. 健康危険情報

特になし

G. 研究発表

1. 論文発表

Nishisako S, Higashi T, Wakao F. Development of AI Chatbots for Cancer Information: Reducing Hallucinations and Trade-Offs in Responses with Reliable Data. JMIR Preprints. 17/12/2024:70176. DOI: 10.2196/preprints.70176. URL: https://preprints.jmir.org/preprint/70176

2. 学会発表

西迫 宗大、東 尚弘、若尾 文彦. 参照する情報源を限定したChatGPTによるがん情報提供のハルシネーション排除の可能性 第61回日本癌治療学会学術集会, 福岡市, 2024/10/26

西迫 宗大、東 尚弘、若尾 文彦. 正しい情報源を参照し返答する生成AIによるがんの情報提供～ハルシネーションと返答割合のトレードオフ～ 第9回築地キャンパス若手職員研究発表会, 東京都, 2025/2/20

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

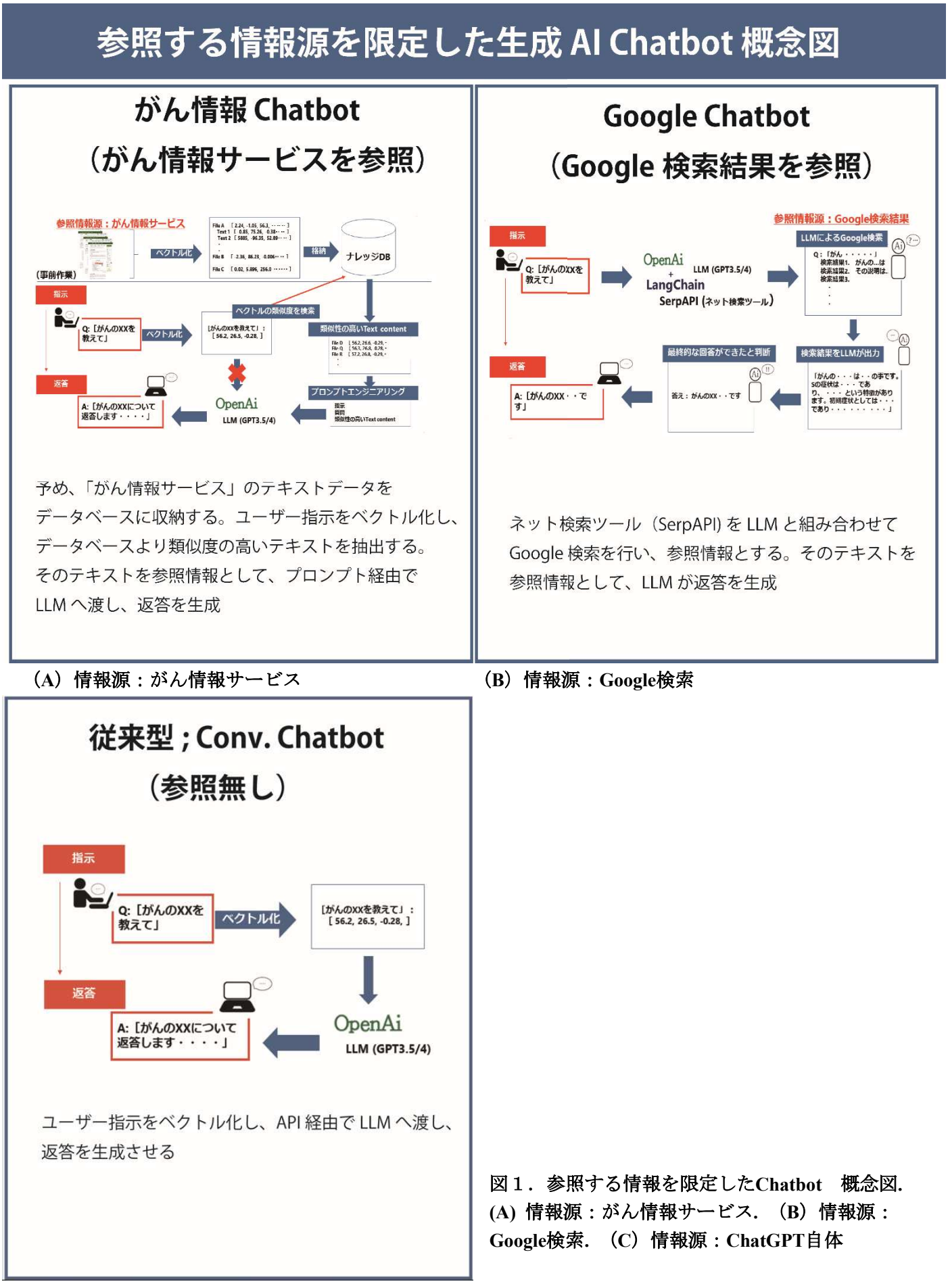

図１．参照する情報を限定したChatbot　概念図.
(A) 情報源：がん情報サービス.（B）情報源：Google検索.（C）情報源：ChatGPT自体

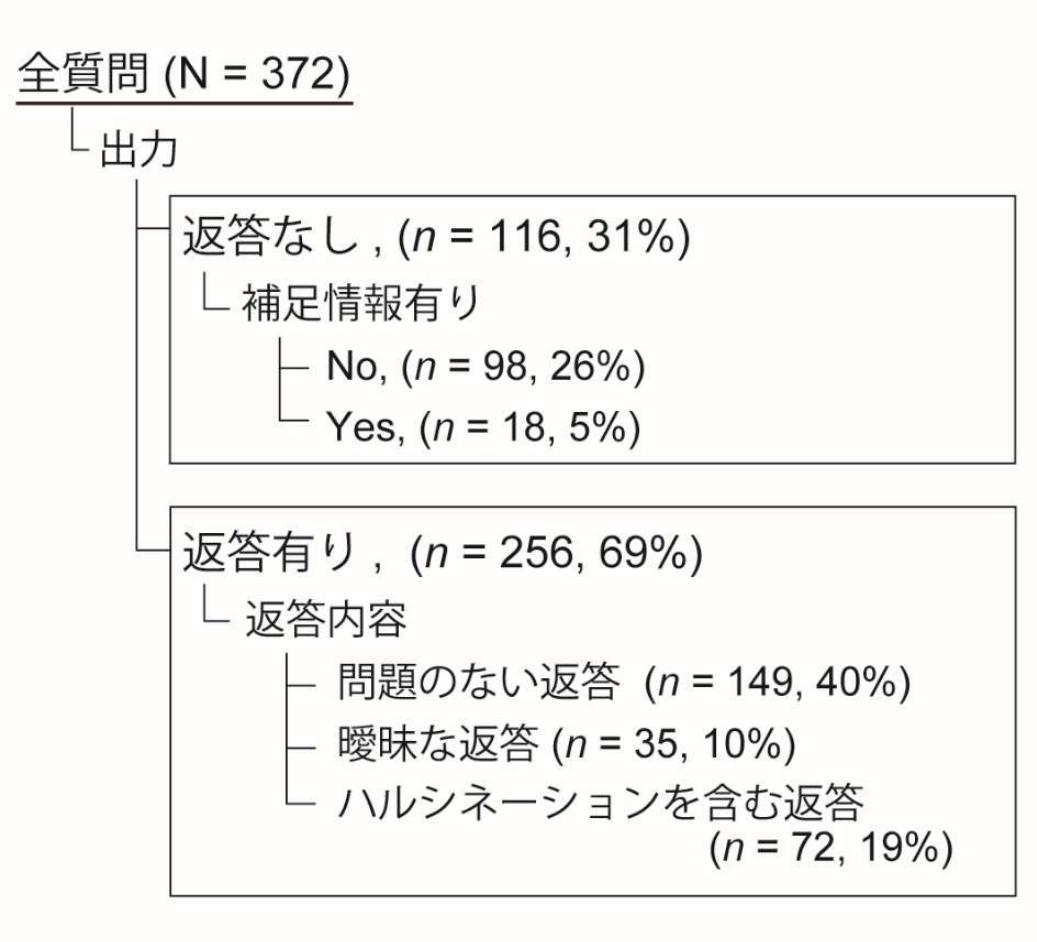

図2. すべての質問に対するAI Chatbot の応答の分類

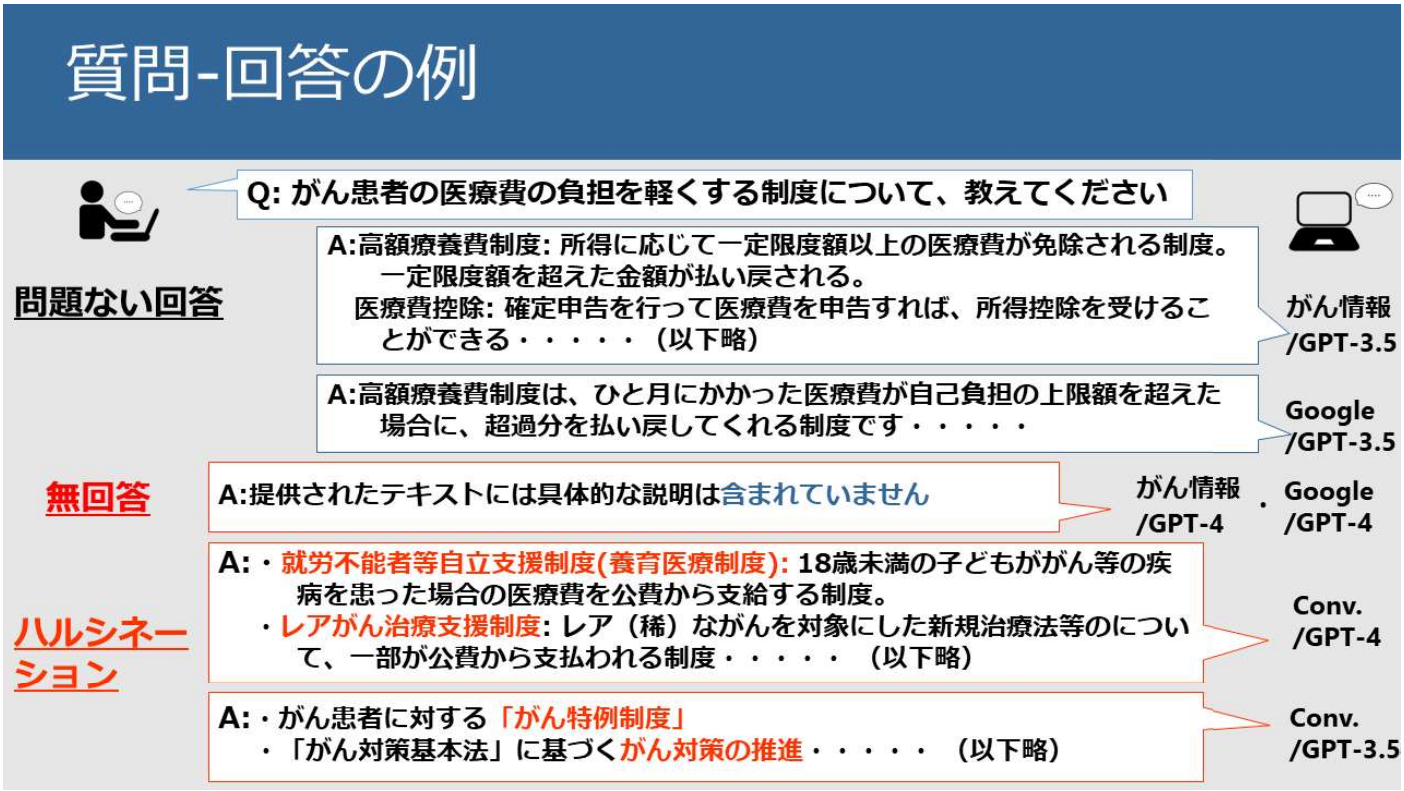

図3. AI-Chatbotの質問に対する応答例とその分類

表1. AI-Chatbotの質問に対する応答の有無

表2. AI-Chatbotの質問に対するハルシネーションを含む返答の割合

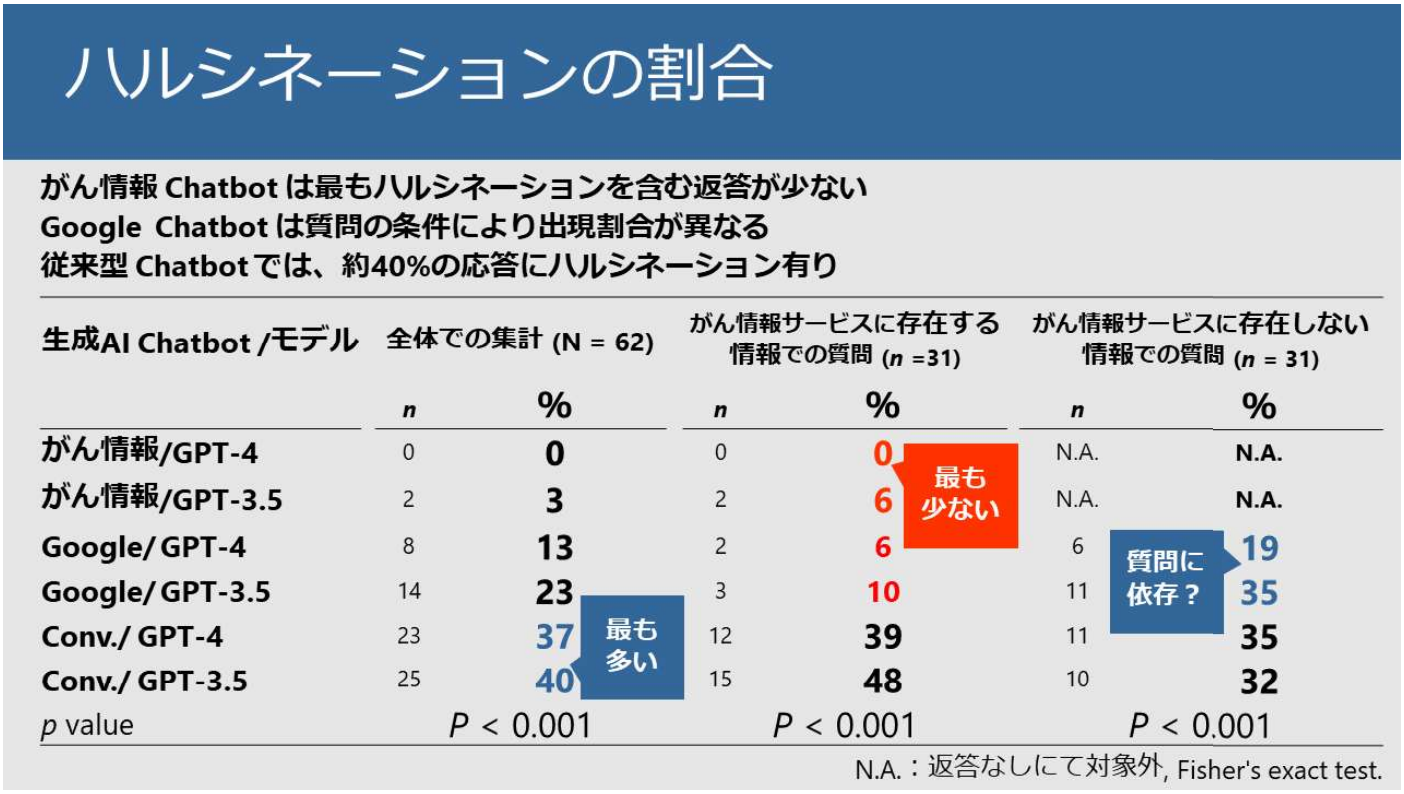

表3. ハルシネーションを含む応答の生成に関連する要因