Original Paper

Evaluating a Chatbot as a Companion for Patients With Breast Cancer: Collaborative Pilot Study

Sebastian Daniel Boie[1], PhD; Esther Glastetter[1], PhD; Michael Patrick Lux[2], MD, MBA; Felix Balzer[3], PhD; Christof von Kalle[4], MD; Christian Lenz[1], MD; Ulrike Müller[1], PhD

[1]Pfizer Pharma GmbH, Berlin, Germany

[2]Department for Gynecology and Obstetrics, St. Louise Women's Hospital, Paderborn, St. Josefs Women's Hospital, Salzkotten, St. Vincenz Clinics, Salzkotten + Paderborn, Germany

[3]Institute of Medical Informatics, Charité - Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[4]Clinical Study Center, Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Berlin, Germany

Corresponding Author:

Sebastian Daniel Boie, PhD
Pfizer Pharma GmbH
Friedrichstr. 110
Berlin 10117
Germany
Phone: 49 15152377580
Email: sebastian.boie@pfizer.com

Abstract

Background: Patients with breast cancer frequently experience significant uncertainty, prompting them to seek detailed, personalized, and reliable medical information to enhance adherence to prescribed treatments, medications, and recommended lifestyle adjustments. Although high-quality information exists within oncology guidelines and patient-oriented resources, the provision of tailored responses to individual patient queries remains challenging, especially for non–English-speaking populations.

Objective: This study aims to evaluate the potential of an artificial intelligence–driven chatbot, specifically leveraging ChatGPT (GPT-4; OpenAI) combined with retrieval-augmented generation, to deliver personalized answers to complex breast cancer-related patient questions in German.

Methods: We collaborated with one of Germany's largest breast cancer Patient Representation Groups to collect authentic patient inquiries, receiving a total of 118 questions. After initial screening, we selected 104 medical questions, organized into 7 distinct categories: aftercare, bone health, ductal carcinoma in situ, diagnostics, nutrition and supplements, complementary medicine, and therapy. A customized version of GPT-4 was configured with specific system prompts emphasizing empathetic, evidence-based responses and integrated with a comprehensive database comprising guidelines, recommendations, and patient information materials published by recognized German medical societies. To assess chatbot responses, we used 4 evaluation criteria: comprehensibility (clarity from a patient perspective), correctness (accuracy per current medical guidelines), completeness (inclusion of all relevant aspects), and potential harm (risk of undue patient harm or misinformation). Ratings were conducted using a 5-point Likert scale by a breast cancer expert (correctness, completeness, and potential harm) and patient representatives (comprehensibility).

Results: The chatbot provided high-quality responses across multiple dimensions. Of the 499 responses evaluated for comprehensibility, 427 (85.6%) were rated as comprehensible. Among the 104 responses assessed for the remaining dimensions, 91 (87.5%) were rated as correct, 72 (69.2%) as complete, and 93 (89.4%) as nonharmful. Reasons for incomplete answers included omission of reimbursement details, updates from recent therapeutic guidelines, or nuanced recommendations regarding endocrine therapy and aftercare schedules. In addition, 6 (5.8%) of the answers were rated as potentially harmful due to outdated or contextually inappropriate recommendations. The chatbot also performed well in the nutrition and bone health categories despite occasionally incomplete document retrieval.

Conclusions: Our findings demonstrate that an artificial intelligence–powered chatbot with GPT-4 and retrieval augmentation can effectively provide personalized, linguistically accessible, and largely accurate information to German-speaking patients

with breast cancer. This approach holds considerable promise for improving patient-centered communication, empowering patients to make informed decisions. Nonetheless, observed limitations regarding response completeness and potential harm underscore the critical need for ongoing human oversight. Future research and development should prioritize regularly updated databases, advanced retrieval methods to handle complex document structures, multimodal capabilities, and clearly articulated disclaimers emphasizing the necessity of professional medical consultation. Our evaluation, along with the provided set of realistic patient questions, establishes a benchmark for future development and validation of German-language oncology chatbots.

JMIR Cancer 2025;11:e68426; doi: 10.2196/68426

Keywords: breast cancer; chatbot; patient information; Generative artificial intelligence; Gen AI; Retrieval-augmented generation; RAG

Introduction

Breast cancer (BC) is the most frequent form of cancer in women [1,2] and a global health concern. Patients with BC have a substantial need for information on their disease at all stages of the patient journey, prefer information tailored to their individual circumstances [3-5], and it is well-known that information on disease and treatments is an important factor for medication adherence and persistence [6-9]. While health care professionals (HCPs) aim to provide comprehensive answers, time constraints and resource limitations can lead to a mismatch in information provision.

Many patients routinely use the internet as a source for information about their condition [5,10], which increases their risk of exposure to misinformation [11,12]. It is expected that patients increasingly turn to artificial intelligence (AI)–based solutions, such as chatbots, which can help to assess the credibility of information. Chatbots can play a significant role in informing patients about their situation and treatment options [13-16]. A substantial benefit of digital information sources is the easy 24-hour access to information and that these sources may answer questions that were left out during consultation by the HCPs. This may help patients have a more informed consultation with their HCP and supports shared decision-making.

Chatbot apps based on large language models (LLMs) are promising as interactive assistants to tailor medical information for patients' specific needs [17-20]. Most existing chatbots are primarily trained on English language data, creating a language barrier for non–English-speaking patients [21,22]. LLM-based chatbots can generate conversational and personalized answers that include context. In general, medical practice differs significantly between different countries due to differences in reimbursements, regulatory frameworks, and cultural attitudes, among other factors [23,24]. While English-speaking LLM-based chatbots are developed for other cancer entities [25], there are, to the best of our knowledge, no existing publications about LLM-powered chatbot solutions for German-speaking patients with BC.

The major disadvantage of LLMs is that they can confidently generate various types of false answers (eg, hallucinations, confabulations, misrepresentations, and omissions among others). A taxonomy of false output and mitigation strategies is a topic of current research [26,27]. A mitigation strategy is to map user questions by classifying their intent to predefined answers. Solutions with predefined answers exist for lung cancer in Japanese [28] and prostate cancer in German [29]. While this approach is generally more reliable, since answers are preselected and quality controlled, it lacks flexibility and personalization of the answers.

Another emerging strategy to mitigate the risk of false information is retrieval-augmented generation (RAG) [30-32] where a trained LLM, such as a generative pretrained transformer (GPT), has access to additional information sources (eg, a database with oncology guidelines and other quality-controlled documents). This additional information can provide up-to-date and quality-controlled context for an LLM to a given query [33].

Ultimately, as AI technologies continue to evolve, health care institutions and organizations may increasingly explore the development of their own LLM-based apps to support more inclusive and patient-centered care. For such efforts to be effective and responsible, they must be grounded in language- and context-specific considerations that reflect patients' real-world concerns.

The aim of this paper is to explore the potential of a retrieval-augmented German-language chatbot based on ChatGPT to address typical information needs of breast cancer patients using real patient questions.

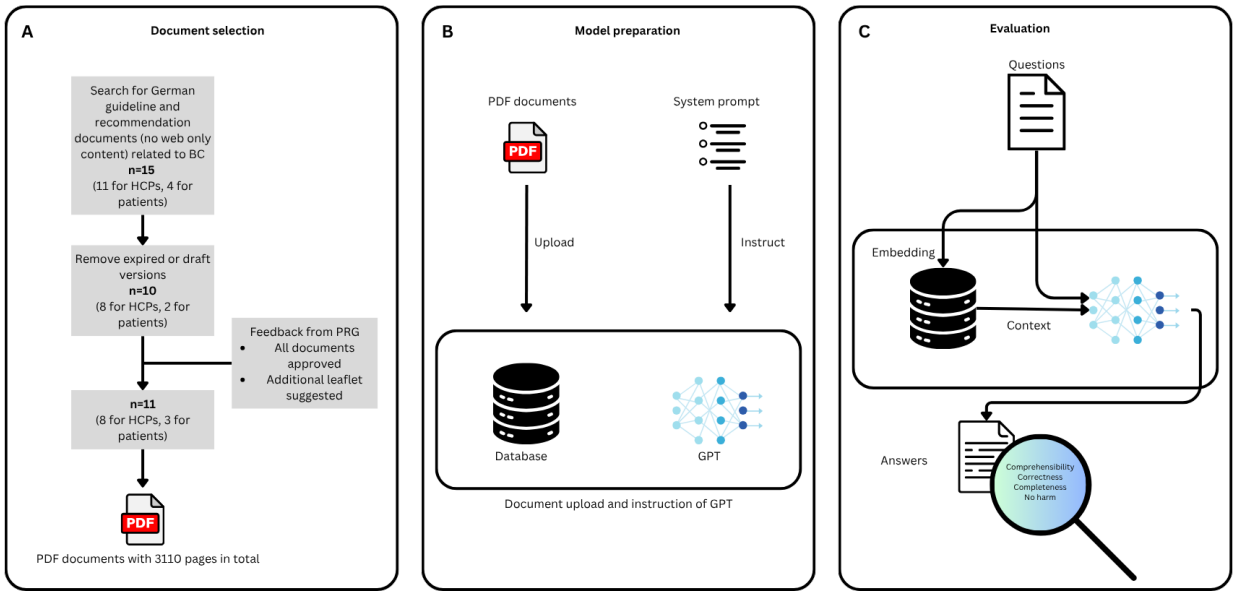
Methods

Overview

This study was conducted in Germany as a collaborative project between researchers from industry and academic experts in digital health, oncology, and AI. An essential partner in this initiative was one of the country's largest breast cancer Patient Representation Groups, comprising individuals with lived experience of breast cancer and deep engagement in patient advocacy. The Patient Representation Group contributed real-world patient questions and helped shape the evaluation criteria, enabling the assessment of the chatbot in a way that reflects the practical needs and concerns of breast cancer patients in the German health care context. This study was performed in accordance with the TRIPOD+LLM (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis+Large Language Model) guidelines (the TRIPOD+LLM checklist is provided in Checklist 1).

The initial phase of the work, including chatbot refinement, question selection, and response evaluation, was completed in April 2024 (see Figure 1).

Figure 1. Overview of the document selection process, chatbot preparation, and evaluation. BC: Breast cancer; HCP: health care professional; PRG: Patient Representation Group.



Document Selection

We selected documents with the purpose of providing the chatbot with up-to-date, evidence-based guidelines and recommendations, ensuring that its responses are grounded in the current standard of care for BC in Germany.

Oncology guidelines and recommendations provided by professional societies are reliable, high-quality sources of information on the diagnosis and treatment of BC. The guidelines are based on expert consensus and compile evidence-based information with a comprehensive coverage and regular updates [34,35]. Guidelines are designed for clinicians to develop tailored diagnosis and treatment plans to the patients' individual tumor biology [36].

Some societies (the German Cancer Society [DKG e.V.], the Arbeitsgemeinschaft Gynäkologische Onkologie e.V. [AGO e.V.], and the Commission Breast of the German Society of Gynecology and Obstetrics) additionally provide information sources directly for patients.

The authors screened and reviewed guidelines and information documents from prominent German medical societies, BC working groups, and nonprofit organizations providing material for HCPs and patients. The entities were selected based on their established role and recognition within the German BC clinical and patient support landscape. A consensus from all authors was reached through discussion and feedback from the Patient Representation Group solicited.

We considered documents from well-known entities (see Textbox 1).

Textbox 1. Medical entities considered for document selection.

We considered documents from the following well-known entities:

- Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e.V. (AWMF).
- Deutsche Krebsgesellschaft e.V. (DKG).
- Deutschen Krebshilfe e.V. (DKH).
- Deutsche Gesellschaft für Gynäkologie und Geburtshilfe e.V. (DGGG).
- PRIO (Prävention und Integrative Onkologie, eine Arbeitsgemeinschaft der DKG).
- Deutsche Gesellschaft für Hämatologie und Medizinische Onkologie e.V. (DGHO).
- Deutschen Gesellschaft für Ernährungsmedizin e.V. (DGEM).
- Arbeitsgemeinschaft, Supportive Maßnahmen in der Onkologie, Rehabilitation und Sozialmedizin der Deutschen Krebsgesellschaft (ASORS).
- Österreichische Arbeitsgemeinschaft für klinische Ernährung (AKE).
- Dachverband der Deutschsprachigen Wissenschaftlichen Osteologischen Gesellschaften e.V. (DVO).
- Arbeitsgemeinschaft Gynäkologische Onkologie e.V. (AGO).
- Arbeitsgemeinschaft für Psychoonkologie in der Deutschen Krebsgesellschaft (PSO).

And selected all documents addressing HCPs or patients that cover diagnosis, treatment, and aftercare of breast cancer; complementary medicine for oncological patients; nutrition in clinical oncology; osteoporosis and psycho-oncological diagnosis, consultation, and treatment.

Each document was manually reviewed, and we excluded documents that are either expired or not yet consented (ie, draft versions and versions for discussion only) from our selection. Draft, discussion, or expired marks (as assigned by the Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e.V.) were verified by 2 authors (SDB and UM). The selection was internally reviewed and discussed with representatives from the BC Patient Representation Group as well as the senior BC expert. All suggestions by one of the authors or the Patient Representation Group made it into the final selection. In total, we included 13 documents comprising 3110 pages (Figure 1A). The documents are available through the respective entities' website, and the compendium can be used by other researchers.

The included documents are shown in Table 1.

A list of excluded documents can be found in Multimedia Appendix 1.

Table 1. Included documents.

Document title	Type of document	Primary target audience	Publisher or leading Medical Societies
Interdisziplinäre S3-Leitlinie für die Früherkennung, Diagnostik, Therapie und Nachsorge des Mammakarzinoms. Langversion 4.4, June 2021, Register 032-045OL	S3 Guideline	HCPs	AWMF, DKG, DKH, DGGG, and DKG.
S3-Leitlinie Komplementärmedizin in der Behandlung von onkologischen PatientInnen Langversion 1.1, September 2021, Register 032/055OL	S3 Guideline	HCPs	AWMF, DKG, DKH, DKG, PRIO, DGGG, and DGHO.
Klinische Ernährung in der Onkologie. 2015, Register 073/006. (DOI 10.1055/s-0035-1552741).	S3 Guideline	HCPs	DGEM with DGHO, ASORS, and AKE.
Prophylaxe, Diagnostik und Therapie der OSTEOPOROSE. Langfassung, September 2023, Register 183/001	S3 Guideline	HCPs	DVO.
Diagnostik und Therapie früher und fortgeschrittener Mammakarzinome 2023.1	Recommendations	HCPs	AGO Breast Commission (of DGGG) and DKG.
Mammakarzinom der Frau. January 2018	Recommendations	HCPs	Onkopedia, DGHO.
Psychoonkologische Diagnostik, Beratung und Behandlung von erwachsenen Krebspatient*innen. Version 2.1 – August 2023	S3 Guideline	HCPs	AWMF, DKG, DKH, DKG, and PSO.
Peri- und Postmenopause – Diagnostik und Interventionen. Register 015-–062, Version 1.1, January 2020	S3 Guideline	HCPs	DGGG.
Patientinnenleitlinie. Brustkrebs im frühen Stadium. December 2018	Patient Guideline based on S3 Guideline	Patients	AWMF, DKG, and Stiftung Deutsche Krebshilfe.
BRUSTKREBS Patientenratgeber zu den AGO-Empfehlungen 2023	Patient companion based on AGO recommendations	Patients	AGO Breast Commission with AGO Patient Forum.
Voiß P. (2018) Möglichkeiten und Grenzen der Komplementärmedizin.	Information leaflet	Patients	Brustkrebs Deutschland e.V.

Model Preparation

We used OpenAI's feature to build custom GPTs, based on GPT-4, with user-defined instructions and access to a document database for RAG [37,38]. The retrieval mechanism can find relevant information from the uploaded documents and pass it along with the original question to the GPT. For our experiments, it was sufficient to upload the documents one by one. All technical details (eg, splitting documents into chunks, embedding chunks to obtain a vector index, question embedding, and similarity-based matching) are handled by OpenAI. The RAG mechanism is used as is, since parameters that affect retrieval (eg, threshold values for similarity measures or number of retrieved documents) are not exposed. A detailed explanation of the RAG technology can be found elsewhere [33,39,40].

We experimented with different instructions in the system prompt using a set of 5 short questions (included in Multimedia Appendix 1) that the authors formulated before receiving test questions from the Patient Representation Group. Based on the initial experiments, we agreed on using the following five instructions: (1) search for relevant information in the documents uploaded; (2) clearly advise against therapies that are not evidence-based; (3) ask clarifying questions, if necessary; (4) formulate empathetic answers; and (5) not to mention severe complications unless they are clearly indicated by the patient; implemented through the

system prompt (Figure 1B). Since our goal is to evaluate the answering capability in the German language, we used the instructions in German (also included in Multimedia Appendix 2). All evaluations took place on or before April 19, 2024.

Test Questions

We asked one of Germany's largest BC Patient Representation Groups to share a set of commonly asked questions. We indicated that we have a focus on medical questions only and made no further recommendations as to topic, number, difficulty, or length of the question. Questions were submitted to the Patient Representation Group in person at regional or national meetings, via phone, email, or various online social network platforms that the BC Patient Representation Group is active on. The Patient Representation Group selected the questions based on their judgment of their importance and frequency of occurrence in real-world patient interactions. They grouped the questions into 7 categories (Aftercare, Bone health, DCIS [ductal carcinoma in situ], Diagnostics, Diet and nutritional supplements, Complementary medicine, and Therapy). Answers were not provided.

We performed an initial screening of each received question to determine whether its nature is medical and excluding legal and reimbursement questions. Exclusion decisions are based on consensus between 2 reviewers (SB and UM). All questions (included and excluded) are included in the Multimedia Appendix 1.

Most remaining questions are used in the evaluation "as-is," even if the question is complex or with a high potential of misinterpretation. We edited some questions by writing out abbreviations or adding the category at the start of the question. The edits are detailed in Table 2.

Table 2. Edits on the original questions before using them to evaluate our chatbot.

Original	Modification	Translation	Affected questions
AI[a]	Aromatase-inhibitor	Aromatase inhibitor	17, 34, 37, 38, 41, 78, and 104.
Empty	Category placed in front of the question for context (eg, Brustkrebs, hormoneller Brustkrebs, Ernährung und Nahrungsergänzungsmittel, Hitzewallungen und Schweißausbrüche (vasomotorische Symptome))	Breast cancer, hormonal breast cancer, nutrition and dietary supplements, excessive sweating (vasomotor symptoms)	74, 75, 76, 77, 89, 96, and 98.
NEM	Nahrungsergänzungsmittel	Dietary supplements	76, 80, 82, 88, and 102.

[a]AI: aromatase inhibitor.

Evaluation

Evaluating the output of LLMs in medical question-answering systems is a topic of current debate, with consortia developing standards [41]. To date, a standardized evaluation framework is missing. Typically, evaluation criteria are defined at the outset of the study and evaluated on a Likert scale [42-45].

We consented to 4 criteria on which to evaluate the chatbot after feedback from the Patient Representation Group. It was agreed that the Patient Representation Group rate the comprehensibility of the answers ("The answer is clear to me."), a senior BC expert (ML), who is represented in several German guideline commissions and on the board of the German Society for Senology, rated correctness ("The answer presents scientifically correct information."), completeness ("The answer includes all important aspects.") and whether the answer has potential to cause undue harm ("The answer does not cause undue harm."; Figure 1C). For the Patient Representation Group, 5 raters were recruited by the spokesperson during an in-person event of regional leaders. Each rater independently completed the evaluation using the 5-point Likert scale and a "don't know / can't answer" option. Only aggregate response counts per item were collected; no individual-level data were recorded. Direct interaction with individual patient raters was not pursued due to legal and ethical considerations.

The 104 questions are posed to the LLM once. To rate an answer, the raters evaluated each statement on a 5-point Likert scale (5=Strongly agree, 4=Agree, 3=Neutral, 2=Disagree, and 1=Strongly disagree). For each of the given criteria, we present individual ratings per category divided by the total number of ratings for the given category.

Ethics Statement

According to §15 of the Professional Code of the Berlin Medical Association, research based solely on anonymized data is exempt from the requirement for formal ethical approval. In line with this provision, we did not obtain ethics committee approval for this study, as only aggregated, deidentified data were used. However, we recognize that the Ethics Committee's guidance encourages researchers to seek consultation even when data are deidentified or aggregated. Prospective consultation with the ethics committee was not sought.

Participants were invited to complete a paper-based form, and there was no direct interaction between the research team and patients. All communication and data collection were managed by a spokesperson from the Patient Representation Group, who aggregated and anonymized the responses before sharing them with the research team. All data were handled in accordance with applicable data protection regulations and shared in anonymized, aggregated form only.

No direct compensation was provided to individual participants. A modest compensation was provided to the Patient Representation Group for their coordination efforts and data aggregation in accordance with the FSA Code for Collaboration with Patient Organizations specified in the

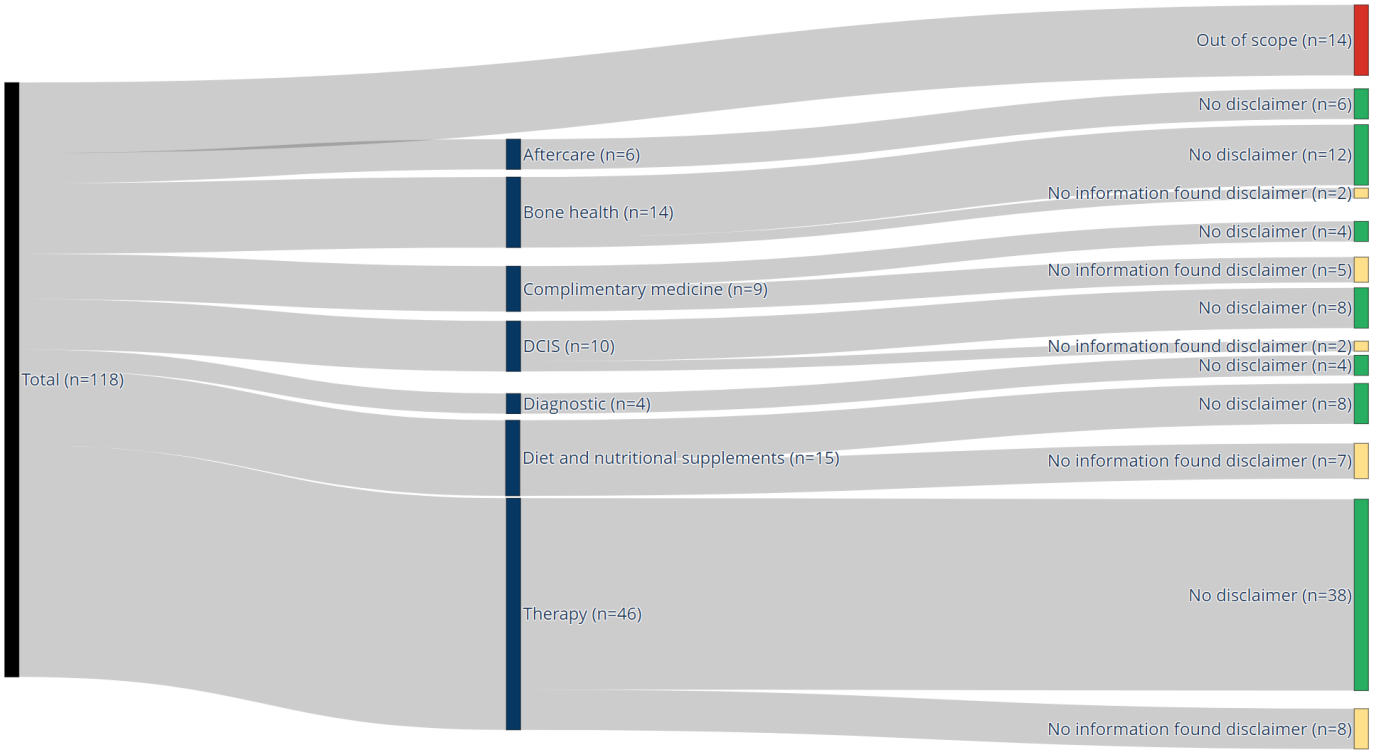
EFPIA Code of Practice (2008). All financial contributions and contracts with Patient Representation Groups are publicly disclosed in the "Transparenzkodex" annually.

Results

Questions

We received a total of 118 questions and a corresponding category for each question. A total of 14 questions are excluded because they were related to nonmedical aspects (eg, insurance coverage, reimbursements, and legal aspects). The questions are grouped into 7 categories (Aftercare, Bone health, DCIS, Diagnostics, Diet and nutritional supplements, Complementary medicine, and Therapy). Figure 2 shows the breakdown of the number of questions per category. Many questions were asked on managing side effects, particularly for endocrine and endocrine-based therapy, with a specific emphasis on complementary and alternative medicine, as well as dietary supplements.

Figure 2. In total, we received 118 questions (black). Human review determined that 14 questions are out of scope (red). The remaining questions are grouped into different categories (blue). Color codes green and yellow indicate whether the model retrieves relevant information from the documents. DCIS: Ductal carcinoma in situ.



For some questions, the LLM reported that no relevant information was found in the documents included. This means that the retriever mechanism was not able to match a query to information stored in the adjacent database. In several cases, the RAG-based model returned no relevant documents from the retrieval component. As internet search was not enabled in our experimental setup, the LLM proceeded to generate responses based solely on its pretrained and fine-tuned knowledge. These instances are reported as part of the results, reflecting the system's behavior when retrieval fails. The model reported for 24 questions that no relevant information was found (see Figure 2).

Chatbot Answers and Evaluation

The answers were evaluated using the four criteria (1) correctness, (2) completeness, (3) no harm, and (4) comprehensibility, either by a senior BC expert (criteria 1-3) or the Patient Representation Group (criteria 4). For each criterion, we formulated a statement (see the Methods section) and evaluated the statement along a 5-point Likert scale. By combining "Strongly Agree" with "Agree" and "Disagree" with "Strongly Disagree," we see that:

- In total, 427 out of 499 (85.6%) ratings of the answers are rated as comprehensible and 42 (8.4%) as incomprehensible.
- A total of 91 out of 104 (87.5%) answers are rated as correct and 7 (6.7%) incorrect.
- In addition, 72 out of 104 (69.2%) answers are rated as including all relevant information, and 20 (19.2%) have some missing information.
- Furthermore, 93 out of 104 (89.4%) answers are rated as not harmful, and 6 (5.7%) may cause potential harm.

The "Neutral" ratings account for the remaining percentages up to 100. Note that comprehensibility was rated by multiple raters, hence the number of ratings exceeds the number of questions.

A total of 6 answers are considered potentially harmful. One answer related to gene expression testing was classified as potentially harmful due to an incorrect statement on reimbursement. In addition, 2 answers on the use

of abemaciclib did not provide information on its use in premenopausal women or failed to mention the absence of data regarding the initiation of abemaciclib therapy 2-3 years after starting endocrine therapy. Furthermore, 2 answers were deemed potentially harmful because they provided individualized recommendations for the duration of aftercare, including breast sonography. Finally, the chatbot mentioned hormone replacement therapy as a treatment for vasomotor symptoms but failed to mention nondrug therapies.

A total of 2 answers were deemed incomplete due to missing reimbursement information. In addition, some answers failed to mention new therapeutic treatment regimens, lacked comprehensive details on aftercare or testing, or omitted aspects related to endocrine therapy, osteo-oncology, or hormonal testing for menopause status.

We analyze the answers for each category separately (see Figure 3 and Table 3). We observe that the chatbot performs well across all criteria for Bone health (14 questions) and diet and nutritional supplements (15 questions). For the latter category, the chatbot reported no relevant information found in the documents for 7 out of 15 questions (see Figure 2). The worst-performing categories (Aftercare and Diagnostics) also have the fewest number of questions (6 and 4, respectively). Some answers (20 out of 104) omit important information, except for the Diet and nutritional supplement category. A total of 6 responses were identified as potentially harmful by expert judgment, typically for omitting important caveats about medication use or not mentioning alternative (non-drug-based) therapies. One answer on abemaciclib therapy did not include information on its use in premenopausal women. Another answer failed to mention the absence of data regarding the initiation of abemaciclib therapy 2-3 years after starting endocrine therapy. Furthermore, 2 answers were deemed potentially harmful because they provided individualized recommendations for the duration and type of aftercare.

Figure 3. Evaluation of the 4 criteria (Comprehensibility, Correctness, Completeness, and No harm) using a 5-point Likert scale for each question category. Compl: Complimentary; DCIS: Ductal carcinoma in situ; Nutr suppl: nutrition supplement.

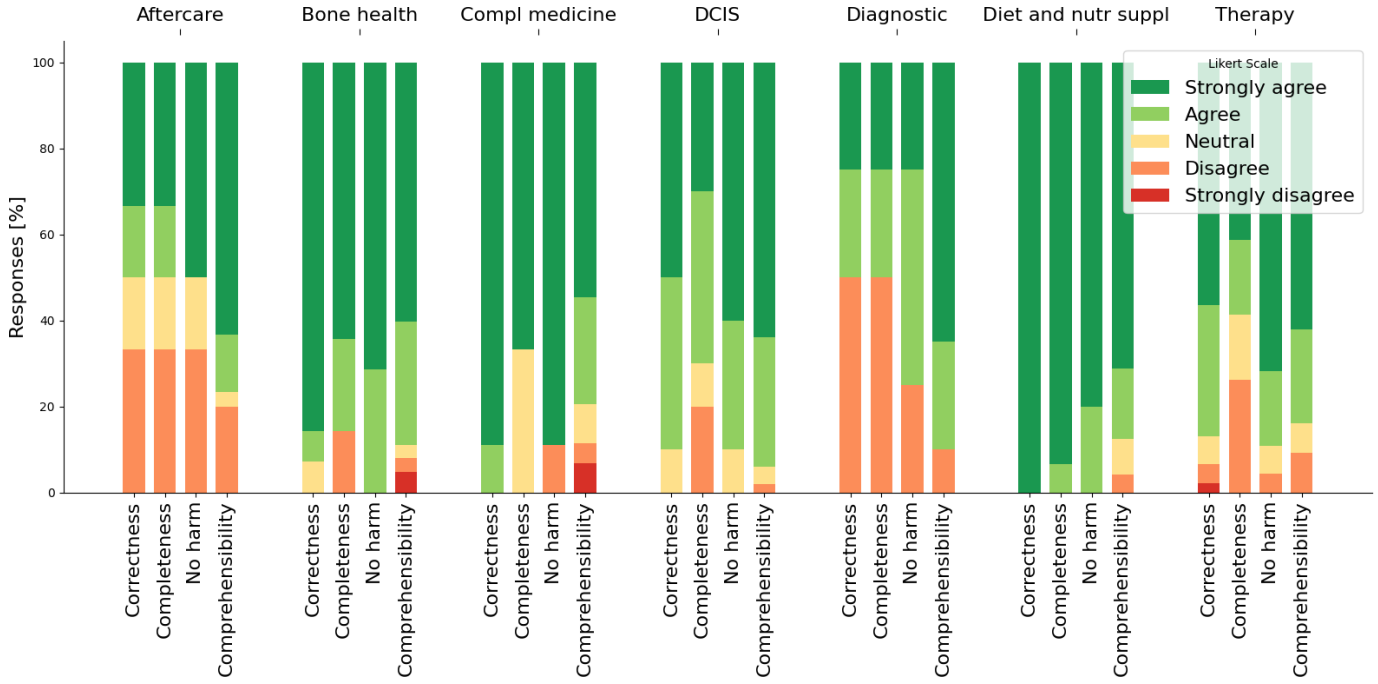


Table 3. Individual ratings per category along the 4 dimensions.

Category	Questions	No information found	Dimension	Ratings, n				
				5	4	3	2	1
Aftercare	6	0	Correctness	2	1	1	2	0
			Completeness	2	1	1	2	0
			No harm	3	0	1	2	0
			Comprehensibility	19	4	1	6	0
Bone health	14	2	Correctness	12	1	1	0	0
			Completeness	9	3	0	2	0
			No harm	10	4	0	0	0
			Comprehensibility	38	18	2	2	3
Complementary medicine	9	5	Correctness	8	1	0	0	0

Category	Questions	No information found	Dimension	Ratings, n				
				5	4	3	2	1
			Completeness	6	0	3	0	0
			No Harm	8	0	0	1	0
			Comprehensibility	24	11	4	2	3
Ductal carcinoma in situ	10	2	Correctness	5	4	1	0	0
			Completeness	3	4	1	2	0
			No harm	6	3	1	0	0
			Comprehensibility	32	15	2	1	0
Diagnostic	4	0	Correctness	1	1	0	2	0
			Completeness	1	1	0	2	0
			No harm	1	2	0	2	0
			Comprehensibility	13	5	0	2	0
Diet and nutritional supplements	15	7	Correctness	15	0	0	0	0
			Completeness	14	1	0	0	0
			No harm	12	3	0	0	0
			Comprehensibility	52	12	6	3	0
Therapy	46	8	Correctness	26	14	3	2	1
			Completeness	19	8	7	12	0
			No harm	33	8	3	2	0
			Comprehensibility	136	48	15	20	0

Finally, the chatbot mentioned hormone replacement therapy as a treatment for vasomotor symptoms (as side effects of endocrine therapy). However, hormone replacement therapy is a contraindication for breast cancer patients in this specific therapeutic situation (Table 3).

Discussion

Principal Findings

One of the largest German BC Patient Representation Group shared a set of typical questions (without identifiable information) that commonly arise. These questions highlight significant and specific information needs of patients—specially regarding endocrine or endocrine-based therapy and the prevention or management of side effects—that go beyond the scope of general health literacy support. Unlike other AI or chatbot or patient vignette studies, this BC Patient Representation Group provided relatively complex medical questions in German, not easily answered through a simple Google search or by referring to patient versions of guidelines.

In response, we evaluated whether a RAG and LLM-based chatbot solution could provide high-quality, tailored information in German. The database included authoritative German-language guidelines and patient information materials that are valid for Germany. Overall, the answers to 104 BC questions were evaluated on 4 criteria (comprehensibility, correctness, completeness, and potential harm) using a 5-point Likert scale. Despite an overall high quality of responses, 20 out of 104 answers were incomplete, and 6 were potentially harmful.

In summary, the chatbot was tested on 104 frequently asked breast cancer–related questions and produced answers that were mostly rated positively across 4 criteria (comprehensibility, correctness, completeness, and harm potential). Specifically, 85.6% (427 out of 499 ratings) of responses were deemed comprehensible, 87.5% (91/104) correct, and 89.4% (93/104) free of undue harm; however, only 69.2% (72/104) of answers were judged complete. Shortcomings primarily involved incomplete details on reimbursement or newer treatment regimens and omissions regarding aftercare or endocrine therapy.

Potentially harmful guidance stemmed mainly from two issues: (1) outdated or ambiguous source material (eg, reimbursement rules that had changed, guidance on hormone replacement therapy lacking caveats about medication use, or not mentioning alternatives) and (2) retrieval gaps that prevented the model from grounding its answer in the most relevant passages. When the necessary nuance was scattered across tables, figures, or inconsistent terminology ("aftercare," "follow-up," and "screening"), the retrieval component sometimes surfaced a partial context, prompting the LLM to fill the vacuum with general knowledge that did not fit the oncological edge cases.

For example, the answer related to gene expression testing was classified as potentially harmful due to an outdated statement on reimbursement found in the source document [46]. In addition, no context was provided for the use of abemaciclib on premenopausal women, likely because the 2018 and 2021 guidelines [47,48] did not provide data on its use in early breast cancer.

The chatbot may have been misled for the recommendation of individualized recommendations for the duration and type of aftercare by the varying terminologies used in the documents ("aftercare," "follow-up," and "screening"), different design formats, tables, and inconsistent time formats (months, years; in numbers or text). In addition, considerations for specific situations such as DCIS, breast scarring after surgery, or the statement that aftercare can be adapted to the symptoms may have contributed to the "confusion." While potentially harmful answers were identified through expert review, actual harm to patients remains unlikely due to clinical safeguards.

Notably, the chatbot performed better than expected in the areas of bone health and nutrition, even when relevant documents were not retrieved, suggesting that its base training data sufficed to answer many diet-related questions. Overall, the system demonstrated promising accuracy and clarity but showed room for improvement in providing comprehensive and fully risk-aware medical guidance.

Several questions encompassed multiple aspects and included subquestions that required broad answers to cover all points (see Multimedia Appendix 1). Incomplete answers were primarily due to the absence of reimbursement details, newer therapeutic regimens not yet documented in guidelines, and retrieval limitations affecting tabular and graphical data representation. In 3 cases, there was a lack of clinical data to address the specific question. For 16 other cases, there was an absence of information in the source documents. Specifically, in 2 of these cases, information on new therapeutic treatment regimens was not yet included in any guidelines. In another case, outdated guideline information led to an inaccurate and incomplete answer. In addition, in 2 cases, the topic of reimbursement was inappropriately covered, despite it typically being out of scope for the source documents. Furthermore, in 8 instances, the available information was not retrieved, probably because it was presented in tables, figures, or listings with ratings (such as on level of evidence and on the grade of recommendation) rather than text, making it accessible only to medical experts who could interpret it, a task beyond the capability of the current retrieval mechanism. The inconsistent terminologies and varying design and time formats across and within the source documents might have also contributed to the nonretrieval of information.

The curation of up-to-date and machine-accessible information in the database improves results. We hypothesize that a textual representation of information from tables, figures, or listings improves results with the given RAG technology. Alternatively, digitized and annotated versions of relevant documents [49] or multimodal RAGs [50] likely improve retrieval.

To mitigate the risk of potentially harmful or incomplete answers, it is important to provide a disclaimer that the chatbot is not able to replace consultation with medical professionals and encourage patients to seek consultation. We strongly suggest that continuous monitoring in the form of transcript reviews and human oversight by medical professionals is implemented, as others have pointed out as well [44]. Furthermore, it is clear that chatbots are classified as medical devices [31] and need substantial safety measures before being used at scale.

Comparison With Previous Work

In agreement with our results, it has been observed that ChatGPT scores high on correctness and often significantly lower on completeness for cancer-related question-answer pairs in an English-language setup [51,52]. A growing body of literature demonstrates that the RAG mechanism helps to ground LLMs in factual information for medical use cases [53,54]. To the best of our knowledge, there is no similar study evaluating LLM-based chatbots on German question-answer pairs for any cancer entity. The training on English-language data may introduce subtle biases from the respective health care systems. Our results show that ChatGPT, together with authoritative documents, can answer BC questions in German with similar performance.

Our realistic set of real-world patient questions can be used by other researchers to develop German BC chatbots. Many questions raised by the BC Patient Representation Group pertain to endocrine and endocrine-based therapy, particularly related to the prevention or management of side effects. Providing active patient support for this therapy is crucial to ensure adherence and prevent early termination, as discontinuation of endocrine or endocrine-based therapy is associated with worse outcomes, including reduced overall survival [55].

A further topic of interest among patients is food supplements and alternative medicine. Evidence-based recommendations on complementary and alternative medicine can help patients avoid negative interactions with cancer treatment, prevent harmful or ineffective therapies, and potentially contribute to treatment success [56]. The chatbot might provide substantial support for patients, especially when based on evidence-based information such as from the German S3 guidelines on complementary medicine in oncology.

This evaluation of a chatbot prototype in German language provides a strong baseline for evaluating German BC chatbot apps. We observe that ChatGPT, together with appropriate documents in the vector store, provides a strong performance even for a German medical question answering task.

Limitations

First, we pose each question to the chatbot only once. Thus, we do not assess whether the retrieval mechanism consistently retrieves the same result or if the LLM's answers are consistent across repeated queries.

Second, the evaluation criteria—correctness, completeness, and potential for undue harm—are only evaluated by a single expert, introducing subjective judgment and potential bias representing a subjective evaluation.

Third, it is possible that the Patient Representation Group introduces a bias in the question selection.

Fourth, our study does not simulate real-world interactions, where patients typically ask follow-up questions and discuss outcomes directly with their treating physicians. A comprehensive real-world evaluation would require extended and interactive dialogues between patients and the chatbot.

Fifth, although we share test questions, source documents and prompts, reproducibility is only partially achievable because our study relies on closed-source algorithms for information retrieval and output generation by the LLM, whose parameters and implementations may change over time.

Future Directions

Since we observe that some questions are not directly addressed in the included documents, a future model can be improved by including relevant journal articles or guidelines and recommendations from other countries. It is well-known that the prompting strategy can influence model performance [57]; therefore, improved prompting (eg, chain-of-thought [58], multiround iterative questioning [59], or self-reasoning [60]) may significantly improve results. In addition, specific fine-tuning on a BC question answering task can be expected to improve performance [61].

A total of 14 questions revolved around insurance coverage, reimbursement, and legal aspects of social law, which underscores the need for specific support in these areas. For our study, we consider these topics out of scope, since an interdisciplinary collaboration between legal and social service experts would be required and a different set of authoritative documents. However, it is crucial to address these issues, as a cancer diagnosis poses a high risk of financial problems for patients [55]. An enhanced chatbot or a chatbot specifically designed to address patient questions in these areas would increase.

Conclusion

A cancer diagnosis is a major turning point in life for most, as it is potentially life-threatening and life-changing. We evaluate an AI-based chatbot in answering realistic and challenging patient questions in the German language. The BC chatbot prototype provides largely accurate, comprehensible, and safe answers for German-speaking patients with BC, but incomplete information remains a limitation, particularly concerning reimbursement and newer treatments. The technology may provide real value today, as it can always be used easily and might help to meet patients' needs for information. Further development, testing, and evaluation of chatbots for patients is a multidisciplinary endeavor that should involve patients actively in this process. In addition, the inclusion of guardrails and prominent disclaimers that technology cannot replace professional consultation and human oversight is important for apps.

Acknowledgments

We would like to thank mamazone–Frauen und Forschung gegen Brustkrebs e. V. for their invaluable support that made this project possible. In addition, we would like to thank Christina Claußen and her Patient Advocacy Team of Pfizer Germany for their many contributions.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

SB, EG, CL, and UM conceptualized the study. SB and UM curated the data and wrote the original draft. ML, FB, and CvK advised on methodology, investigation, and formal analysis. All authors read and approved the final manuscript.

Conflicts of Interest

MPL was a paid consultant to Pfizer in connection with the development of this paper. FB was a paid consultant to Pfizer in connection with the development of this paper. CvK was a paid consultant to Pfizer in connection with the development of this paper. SDB, EG, CL, and UM are salaried employees at Pfizer Pharma GmbH and shareholders of Pfizer.

Multimedia Appendix 1

Questions for initial experimentation, prompts, and incomplete answers.
[DOCX File (Microsoft Word File), 20 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Questions of the Patient Representation Group along with answers by the chatbot.
[XLSX File (Microsoft Excel File), 77 KB-Multimedia Appendix 2]

Checklist 1

TRIPOD+LLM (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis+Large Language Model) checklist.
[PDF File (Adobe File), 176 KB-Checklist 1]

References

1. Arnold M, Morgan E, Rumgay H, et al. Current and future burden of breast cancer: global statistics for 2020 and 2040. Breast. Dec 2022;66:15-23. [doi: 10.1016/j.breast.2022.08.010] [Medline: 36084384]
2. Łukasiewicz S, Czeczelewski M, Forma A, Baj J, Sitarz R, Stanisławek A. Breast cancer-epidemiology, risk factors, classification, prognostic markers, and current treatment strategies-an updated review. Cancers (Basel). Aug 25, 2021;13(17):4287. [doi: 10.3390/cancers13174287] [Medline: 34503097]
3. Mistry A, Wilson S, Priestman T, Damery S, Haque M. How do the information needs of cancer patients differ at different stages of the cancer journey? A cross-sectional survey. JRSM Short Rep. Sep 15, 2010;1(4):30. [doi: 10.1258/shorts.2010.010032] [Medline: 21103122]
4. Tran Y, Lamprell K, Nic Giolla Easpaig B, Arnolda G, Braithwaite J. What information do patients want across their cancer journeys? A network analysis of cancer patients' information needs. Cancer Med. Jan 2019;8(1):155-164. [doi: 10.1002/cam4.1915] [Medline: 30525298]
5. van Eenbergen M, Vromans RD, Boll D, et al. Changes in internet use and wishes of cancer survivors: A comparison between 2005 and 2017. Cancer. Jan 15, 2020;126(2):408-415. [doi: 10.1002/cncr.32524] [Medline: 31580497]
6. Buyens G, van Balken M, Oliver K, et al. Cancer literacy - Informing patients and implementing shared decision making. J Cancer Policy. Mar 2023;35:100375. [doi: 10.1016/j.jcpo.2022.100375] [Medline: 36462750]
7. Heisig SR, Shedden-Mora MC, von Blanckenburg P, et al. Informing women with breast cancer about endocrine therapy: effects on knowledge and adherence. Psychooncology. Feb 2015;24(2):130-137. [doi: 10.1002/pon.3611] [Medline: 24953538]
8. Kvarnström K, Westerholm A, Airaksinen M, Liira H. Factors contributing to medication adherence in patients with a chronic condition: a scoping review of qualitative research. Pharmaceutics. Jul 20, 2021;13(7):1100. [doi: 10.3390/pharmaceutics13071100] [Medline: 34371791]
9. Verma S, Madarnas Y, Sehdev S, Martin G, Bajcar J. Patient adherence to aromatase inhibitor treatment in the adjuvant setting. Curr Oncol. May 2011;18 Suppl 1(Suppl 1):S3-9. [doi: 10.3747/co.v18i0.899] [Medline: 21698059]
10. Lu X, Zhang R, Wu W, Shang X, Liu M. Relationship between internet health information and patient compliance based on trust: empirical study. J Med Internet Res. Aug 17, 2018;20(8):e253. [doi: 10.2196/jmir.9364] [Medline: 30120087]
11. Chen L, Wang X, Peng TQ. Nature and diffusion of gynecologic cancer–related misinformation on social media: analysis of tweets. J Med Internet Res. Oct 16, 2018;20(10):e11515. [doi: 10.2196/11515] [Medline: 30327289]
12. Lazard AJ, Nicolla S, Vereen RN, et al. Exposure and reactions to cancer treatment misinformation and advice: survey study. JMIR Cancer. Jul 28, 2023;9:e43749. [doi: 10.2196/43749] [Medline: 37505790]
13. Fritsch SJ, Blankenheim A, Wahl A, et al. Attitudes and perception of artificial intelligence in healthcare: a cross-sectional survey among patients. Digit Health. 2022;8:20552076221116772. [doi: 10.1177/20552076221116772] [Medline: 35983102]
14. Hopkins AM, Logan JM, Kichenadasse G, Sorich MJ. Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift. JNCI Cancer Spectr. Mar 1, 2023;7(2):pkad010. [doi: 10.1093/jncics/pkad010] [Medline: 36808255]
15. Hudecek MFC, Lermer E, Gaube S, Cecil J, Heiss SF, Batz F. Fine for others but not for me: the role of perspective in patients' perception of artificial intelligence in online medical platforms. Computers in Human Behavior: Artificial Humans. Jan 2024;2(1):100046. [doi: 10.1016/j.chbah.2024.100046]
16. Young AT, Amara D, Bhattacharya A, Wei ML. Patient and general public attitudes towards clinical artificial intelligence: a mixed methods systematic review. Lancet Digit Health. Sep 2021;3(9):e599-e611. [doi: 10.1016/S2589-7500(21)00132-1] [Medline: 34446266]
17. Bitterman DS, Downing A, Maués J, Lustberg M. Promise and perils of large language models for cancer survivorship and supportive care. J Clin Oncol. May 10, 2024;42(14):1607-1611. [doi: 10.1200/JCO.23.02439] [Medline: 38452323]
18. Haase I, Xiong T, Rissmann A, Knitza J, Greenfield J, Krusche M. ChatSLE: consulting ChatGPT-4 for 100 frequently asked lupus questions. Lancet Rheumatol. Apr 2024;6(4):e196-e199. [doi: 10.1016/S2665-9913(24)00056-0] [Medline: 38508817]
19. Siglen E, Vetti HH, Lunde ABF, et al. Ask Rosa - The making of a digital genetic conversation tool, a chatbot, about hereditary breast and ovarian cancer. Patient Educ Couns. Jun 2022;105(6):1488-1494. [doi: 10.1016/j.pec.2021.09.027] [Medline: 34649750]
20. Sun H, Zhang K, Lan W, et al. An AI dietitian for type 2 diabetes mellitus management based on large language and image recognition models: preclinical concept validation study. J Med Internet Res. Nov 9, 2023;25:e51300. [doi: 10.2196/51300] [Medline: 37943581]
21. Al Shamsi H, Almutairi AG, Al Mashrafi S, Al Kalbani T. Implications of language barriers for healthcare: a systematic review. Oman Med J. Mar 2020;35(2):e122. [doi: 10.5001/omj.2020.40] [Medline: 32411417]

22. Bressem KK, Papaioannou JM, Grundmann P, et al. medBERT.de: a comprehensive German BERT model for the medical domain. Expert Syst Appl. Mar 2024;237:121598. [doi: 10.1016/j.eswa.2023.121598]
23. Tikkanen R, Osborn R, Mossialos E, Djordjevic A. International Profiles of Health Care Systems. Commonwealth Fund; 2020. URL: https://www.commonwealthfund.org/sites/default/files/2020-12/International_Profiles_of_Health_Care_Systems_Dec2020.pdf
24. von dem Knesebeck O, Bönte M, Siegrist J, et al. Country differences in the diagnosis and management of coronary heart disease - a comparison between the US, the UK and Germany. BMC Health Serv Res. Sep 29, 2008;8(1):198. [doi: 10.1186/1472-6963-8-198] [Medline: 18823556]
25. Khene ZE, Bigot P, Mathieu R, Rouprêt M, Bensalah K, French Committee of Urologic Oncology. Development of a personalized chat model based on the european association of urology oncology guidelines: harnessing the power of generative artificial intelligence in clinical practice. Eur Urol Oncol. Feb 2024;7(1):160-162. [doi: 10.1016/j.euo.2023.06.009] [Medline: 37474402]
26. Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. arXiv. Preprint posted online on Nov 9, 2023. [doi: 10.48550/arXiv.2311.05232]
27. Rawte V, Chakraborty S, Pathak A, et al. The troubling emergence of hallucination in large language models -- an extensive definition, quantification, and prescriptive remediations. arXiv. Preprint posted online on Oct 8, 2023. [doi: 10.48550/arXiv.2310.04988]
28. Kataoka Y, Takemura T, Sasajima M, Katoh N. Development andearly feasibility of chatbots for educating patients with lung cancer and their caregivers in Japan: mixed methods study. JMIR Cancer. Mar 10, 2021;7(1):e26911. [doi: 10.2196/26911] [Medline: 33688839]
29. Görtz M, Baumgärtner K, Schmid T, et al. An artificial intelligence-based chatbot for prostate cancer education: Design and patient evaluation study. Digit Health. 2023;9:20552076231173304. [doi: 10.1177/20552076231173304] [Medline: 37152238]
30. Ge J, Sun S, Owens J, et al. Development of a liver disease-specific large language model chat interface using retrieval-augmented generation. Hepatology. Nov 1, 2024;80(5):1158-1168. [doi: 10.1097/HEP.0000000000000834] [Medline: 38451962]
31. Gilbert S, Harvey H, Melvin T, Vollebregt E, Wicks P. Large language model AI chatbots require approval as medical devices. Nat Med. Oct 2023;29(10):2396-2398. [doi: 10.1038/s41591-023-02412-6] [Medline: 37391665]
32. Tonmoy S, Zaman SMM, Jain V, et al. A comprehensive survey of hallucination mitigation techniques in large language models. Jan 2, 2024. [doi: 10.48550/arXiv.2401.01313]
33. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. May 22, 2020. [doi: 10.48550/arXiv.2005.11401]
34. Reames BN, Krell RW, Ponto SN, Wong SL. Critical evaluation of oncology clinical practice guidelines. J Clin Oncol. Jul 10, 2013;31(20):2563-2568. [doi: 10.1200/JCO.2012.46.8371] [Medline: 23752105]
35. Untch M, Fasching PA, Brucker SY, et al. Behandlung von Patientinnen mit frühem Mammakarzinom: Evidenz, Kontroversen, Konsens – Meinungsbild deutscher Expert*innen zur 17. Internationalen St.-Gallen-Konsensuskonferenz [Article in German]. Senologie - Zeitschrift Für Mammadiagnostik und -therapie. Jun 2021;18(2):163-181. [doi: 10.1055/a-1463-8544]
36. Gradishar WJ, Anderson BO, Abraham J, et al. Breast cancer, version 3.2020, NCCN clinical practice guidelines in oncology. J Natl Compr Canc Netw. Apr 2020;18(4):452-478. [doi: 10.6004/jnccn.2020.0016] [Medline: 32259783]
37. OpenAI, Adler S, Agarwal S, et al. GPT-4 technical report. arXiv. Preprint posted online on Mar 15, 2024. [doi: 10.48550/arXiv.2303.08774]
38. OpenAI. Introducing GPTs. 2023. URL: https://openai.com/index/introducing-gpts/
39. Finardi P, Avila L, Castaldoni R, et al. The chronicles of RAG: the retriever, the chunk and the generator. 2024. URL: http://arxiv.org/abs/2401.07883
40. Popat SK, Deshmukh PB, Metre VA. Hierarchical document clustering based on cosine similarity measure. Presented at: 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM). Oct 5-6, 2017:IEEE. 153-159; Aurangabad, India. [doi: 10.1109/ICISIM.2017.8122166]
41. The CHART Collaborative. Protocol for the development of the Chatbot Assessment Reporting Tool (CHART) for clinical advice. BMJ Open. May 2024;14(5):e081155. [doi: 10.1136/bmjopen-2023-081155]
42. 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. Aug 1, 2023;6(8):e2330320. [doi: 10.1001/jamanetworkopen.2023.30320] [Medline: 37606922]

43. Jeblick K, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol. May 2024;34(5):2817-2825. [doi: 10.1007/s00330-023-10213-1] [Medline: 37794249]
44. Maroncelli R, Rizzo V, Pasculli M, et al. Probing clarity: AI-generated simplified breast imaging reports for enhanced patient comprehension powered by ChatGPT-4o. Eur Radiol Exp. Oct 30, 2024;8(1):124. [doi: 10.1186/s41747-024-00526-1] [Medline: 39477904]
45. Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol. Jul 2023;29(3):721-732. [doi: 10.3350/cmh.2023.0089] [Medline: 36946005]
46. Janni W, Müller V. Patientenratgeber Zu Den AGO-Empfehlungen 2023 [Book in German]. Zuckschwerdt Verlag München; 2023.
47. Wöckel A, Kreienberg R, Janni W. Interdisziplinäre S3-leitlinie für die früherkennung, diagnostik. In: Therapie Und Nachsorge Des Mammakarzinoms [Book in German]. 2021.
48. Wörmann B, Aebi S, Balic M, et al. Mammakarzinom der frau. In: DGHO Deutsche Gesellschaft Für Hämatologie Und Medizinische Onkologie eV [Book in German]. 2018.
49. Borchert F, Lohr C, Modersohn L, et al. GGPONC: a corpus of german medical text with rich metadata based on clinical practice guidelines. arXiv. Preprint posted online on Jul 13, 2020. [doi: 10.48550/arXiv.2007.06400]
50. Lahiri AK, Hu QV. AlzheimerRAG: multimodal retrieval augmented generation for PubMed articles. Dec 21, 2024. [doi: 10.48550/arXiv.2412.16701]
51. Goodman RS, Patrinely JR, Stone CA Jr, et al. Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open. Oct 2, 2023;6(10):e2336483. [doi: 10.1001/jamanetworkopen.2023.36483] [Medline: 37782499]
52. Iannantuono GM, Bracken-Clarke D, Floudas CS, Roselli M, Gulley JL, Karzai F. Applications of large language models in cancer care: current evidence and future perspectives. Front Oncol. 2023;13:1268915. [doi: 10.3389/fonc.2023.1268915] [Medline: 37731643]
53. Ng KKY, Matsuba I, Zhang PC. RAG in health care: a novel framework for improving communication and decision-making by addressing LLM limitations. NEJM AI. Jan 2025;2(1). [doi: 10.1056/AIra2400380]
54. Quidwai MA, Lagana A. A RAG chatbot for precision medicine of multiple myeloma. Genetic and Genomic Medicine. Preprint posted online on 2024. [doi: 10.1101/2024.03.14.24304293]
55. Eliassen FM, Blåfjelldal V, Helland T, et al. Importance of endocrine treatment adherence and persistence in breast cancer survivorship: a systematic review. BMC Cancer. Jul 4, 2023;23(1):625. [doi: 10.1186/s12885-023-11122-8] [Medline: 37403065]
56. Johnson SB, Park HS, Gross CP, Yu JB. Complementary medicine, refusal of conventional cancer therapy, and survival among patients with curable cancers. JAMA Oncol. Oct 1, 2018;4(10):1375-1381. [doi: 10.1001/jamaoncol.2018.2487] [Medline: 30027204]
57. Sivarajkumar S, Kelley M, Samolyk-Mazzanti A, Visweswaran S, Wang Y. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study. JMIR Med Inform. Apr 8, 2024;12:e55318. [doi: 10.2196/55318] [Medline: 38587879]
58. Lee GG, Latif E, Wu X, Liu N, Zhai X. Applying large language models and chain-of-thought for automatic scoring. Computers and Education: Artificial Intelligence. Jun 2024;6:100213. [doi: 10.1016/j.caeai.2024.100213]
59. Yuan J, Bao P, Chen Z, et al. Advanced prompting as a catalyst: empowering large language models in the management of gastrointestinal cancers.
60. Xia Y, Zhou J, Shi Z, Chen J, Huang H. Improving retrieval augmented language model with self-reasoning. Jul 29, 2024. [doi: 10.48550/arXiv.2407.19813]
61. Nguyen Z, Annunziata A, Luong V, et al. Enhancing Q&A with domain-specific fine-tuning and iterative reasoning: a comparative study. arXiv. Apr 17, 2024. [doi: 10.48550/arXiv.2404.11792]

Abbreviations

AGO: Arbeitsgemeinschaft Gynäkologische Onkologie e.V.
AI: artificial intelligence
BC: breast cancer
DCIS: ductal carcinoma in situ
GPT: Generative pretrained transformer
HCP: health care professional
LLM: large language model
RAG: retrieval-augmented generation

TRIPOD+LLM: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis+Large Language Model

Edited by Naomi Cahill; peer-reviewed by Chenxu Wang, Minjin Kim, Veronica Rizzo; submitted 06.11.2024; final revised version received 13.06.2025; accepted 17.06.2025; published 13.08.2025

Please cite as:
Boie SD, Glastetter E, Lux MP, Balzer F, von Kalle C, Lenz C, Müller U
Evaluating a Chatbot as a Companion for Patients With Breast Cancer: Collaborative Pilot Study
JMIR Cancer 2025;11:e68426
URL: https://cancer.jmir.org/2025/1/e68426
doi: 10.2196/68426