Improved precision oncology question-answering using agentic LLM

Rangan Das[5,§], K Maheswari[1,§], Shaheen Siddiqui[1,*], Nikita Arora[1,*], Ankush Paul[1,*], Jeet Nanshi[1,*], Varun Udbalkar[1], Apoorva Sarvade[1], Harsha Chaturvedi[1], Tammy Shvartsman[1], Shet Masih[1], R Thippeswamy[6], Shekar Patil[6], S S Nirni[7], Brian Garsson[1], Sanghamitra Bandyopadhyay[8], Ujjwal Maulik[1,5], Mohammed Farooq[1], Debarka Sengupta[1,2,3,4+]

1. GeneSilico, Inc., 3267 Bee Caves Rd, STE 107-332 Austin, TX 78746
2. Department of Computational Biology, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), Okhla, Phase III, New Delhi, 110020, India
3. Department of Computer Science and Engineering, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), Okhla, Phase III, New Delhi, 110020, India
4. Centre for Artificial Intelligence, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), Okhla, Phase III, New Delhi, 110020, India
5. Department of Computer Science and Engineering, Jadavpur University, Kolkata, India
6. Department of Medical Oncology, HGC Cancer Centre, Bangalore, Karnataka 560027, India
7. Department of Medical Oncology, Omega Hospitals and Indo-American Cancer Institute and Research Centre, Hyderabad, Andhra Pradesh, India
8. Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, India

+ To whom correspondence should be addressed:

dsengupta@genesilico.ai

§ co-first authors, * equal contribution

ABSTRACT

Despite the widespread application of Large Language Models (LLMs) in biomedical research, their clinical adoption faces significant challenges. These challenges stem from concerns about the quality, accuracy, and comprehensiveness of LLM-generated answers. Most existing work has focused on fine-tuning LLMs based on foundation models, which have not yet fully addressed accuracy and reliability issues. In this work, we propose an agent-based approach that aims to make LLM-based systems clinically deployable for precision oncology, while mitigating common pitfalls such as hallucinations, incoherence, and "lost-in-the-middle" problems. To achieve this, we implemented an agentic architecture, fundamentally shifting an LLM's role from a simple response synthesizer to planner. This agent orchestrates a suite of specialized tools that asynchronously retrieve information from various sources. These tools include curated document vector stores encompassing treatment guidelines, genomic data, clinical trial information, drug data, and breast cancer literature. The LLM then leverages its planning capabilities to synthesize information retrieved by these tools, generating comprehensive and accurate responses. We demonstrate GeneSilico Copilot's effectiveness in the domain of breast cancer, achieving state-of-the-art accuracy. Furthermore, the system showcases success in generating personalized oncotherapy recommendations for real-world cases.

INTRODUCTION

Cancer's inherent complexity, driven by both inter and intra-tumoral heterogeneity, presents a significant hurdle in clinical management. This genetic heterogeneity allows tumors to evade traditional treatments. However, the advent of precision oncology has led to the development of genome-targeted and genome-informed therapies, which aim to address this challenge. From 2006 to 2020, the eligibility for genome-targeted therapies in the U.S. increased from 5.13% to 13.60%, while the response rate improved from 2.73% to 7.04%. Similarly, genome-informed therapies saw a rise in eligibility from 10.70% to 27.30% and an increase in response from 3.33% to 11.10% during the same period. Interestingly, most of the eligibility increase for genome-targeted therapies occurred after 2018, whereas most of the response increase was observed before 2018. These findings highlight a concerning trend: while eligibility for these therapies is on the rise, the actual response rate remains low[1]. Large Language Models (LLMs) offer a powerful opportunity to address these challenges[2]. By leveraging their ability to process and synthesize vast amounts of healthcare data, LLMs can assist oncologists in navigating the ever-expanding landscape of targeted therapies. They can analyze a patient's specific genetic profile, identify relevant clinical trials and treatment guidelines, and even suggest potential drug combinations tailored to the unique characteristics of the patient's cancer. While LLMs can correctly identify some key strategies and offer reasonable, albeit incomplete, suggestions even experts missed, they can also generate factual errors (hallucination), irrelevant, harmful, or biased content[3].

While multiple industries have adopted LLM-based models, adoption in biology and medicine is still lacking[4]. With the ever-evolving treatment guidelines and drug approval[5,6] it is difficult for fine-tuned LLMs to stay updated. State of the art models like Med-PaLM [7], BioGPT [8], and BioBERT [9] are standalone systems that are trained and fine-tuned on domain-specific large-scale biomedical corpora. These have achieved notable results on medical question answering datasets. GatorTron[10] is a similar model that has been designed and trained from scratch, and subsequently fine-tuned for tasks like clinical concept extraction, medical relation extraction, semantic textual similarity, natural language inference, and medical question answering. While there are domain-specific LLMs, they come with the same drawbacks of foundational

models. Foundational models such as ChatGPT 3.5 itself can act as a support tool for breast tumor board, but suffer from the lack of references, or the potential to produce seemingly credible but incorrect responses[11]. In the domain of radiation oncology, ChatGPT 3.5 showed high accuracy and completeness in radiation oncology queries, but higher-than-recommended readability levels suggest the need for refinement for improved patient accessibility and understanding[12]. Newer models such as GPT-4 and models from Anthropic provide better responses, but all LLMs continue to have clinically significant error rates, including examples of overconfidence and consistent inaccuracies[13]. In the context of treatment guidelines, ChatGPT provides concise, accessible supportive care advice including many non-medical support recommendations, but its recommendations lacked the specificity observed in National Comprehensive Cancer Network (NCCN) guidelines including often not suggesting any medications[14]. These discrepancies with guidelines raise concerns for patient-facing symptom management recommendations. This can be partly attributed to the fact that these models are trained on publicly available data and do not have sufficient specialized domain information. Models implemented on more focused domains have better performance. CancerBERT[15] is an example of a model trained on a narrower domain, but the model has only been evaluated on named entity recognition (NER) tasks. The output of standalone models can be further optimized through retrieval augmented generation (RAG). In RAGs, the LLM retrieves information from pre-defined storage and synthesizes the response. External data can be used to augment the response, thereby providing more context and reducing false information. For example, in clinical trial screening, GPT-4 has shown promising performance when augmented with external data sources[16]. Similarly, GPT-4 has also been used for retrieving cancer guidelines. GPT-4 with RAG provided significantly higher correct responses when compared to the standalone LLM service[17]. RefAI[18] is a similar tool that uses retrieval augmented generation to fetch medical literature in real time and summarize them. These examples show the potential of retrieval augmented generation to mitigate the shortcomings of standalone LLM services.

While simple RAG systems demonstrate proficiency in information retrieval and response synthesis, they remain susceptible to issues inherent in the retrieval process. These limitations include the retrieval of

irrelevant information, the failure to capture context during retrieval, and the improper re-ranking of retrieved documents. Agent-based frameworks represent the next step in evolution, enabling LLMs to utilize various tools for information gathering and subsequent response synthesis through their own inherent reasoning capabilities. These tools empower agents to function as multi-tasking systems, with each tool specializing in a specific task. This multi-faceted approach becomes particularly crucial in the domain of precision oncology, where expertise across various verticals is essential.

The GeneSilico Copilot (GSCP) framework exemplifies an agent-based approach designed around a unique hybrid retrieval system. This system integrates a meticulously curated corpus with a novel retrieval approach. The GSCP employs a ReAct agent[19] to generate reasoning traces alongside textual actions. These traces inform the model's internal state, ultimately enhancing its decision-making capabilities. Notably, the final response incorporates these reasoning traces. The model utilizes function calls (tools) to execute a reason-action loop. These tools, implemented as semantic search retrievers, connect to vector database. They process retrieved information by summarizing, refining, or reranking them. Using information from different tools, the agent finally generates the response.

RESULTS

GSCP, an agentic framework for precision oncology

Precision oncology requires in-depth information but focuses on a finite number of key aspects, such as drug dosage information, treatment guidelines, or potential clinical trials. A single retrieval system might not be sufficient to navigate this complex and nuanced landscape. Here, we propose an agent-based approach that empowers the LLM to control and utilize multiple specialized tools based on evolving conversation. These tools can include retrievers, re-rankers, and summarizers, along with hybrid search functionalities. The agent uses its planning and reasoning to synthesize the response by selecting the most appropriate tool for each information need. This allows the agent to, for instance, recommend personalized therapy options by looking up relevant literature using a specific tool, or utilize another tool to suggest clinical trials suitable for the patient's specific condition. Similarly, the agent can recommend drugs and appropriate dosages by employing a tool specializing in drug information retrieval. The outputs from these

tools are then fed back into the LLM, informing its reasoning and synthesis processes, ultimately leading to the generation of a comprehensive and informative response tailored to the user's needs (Figure 1a).

In contrast, traditional RAG systems typically rely on a single retrieval approach, which can be limiting in complex domains like precision oncology. For instance, a simple keyword search might miss relevant documents due to synonymous terminology or variations in phrasing. Our agent-based approach overcomes these limitations by providing greater control and transparency over the retrieval process. The ReAct agent can dynamically select the most suitable tool based on the current conversation state and the user's information requirements. A key benefit of the ReAct agent is its ability to expose the reasoning process behind the synthesized response (Figure 1b).

For data retrieval, the tools in GSCP employ a hybrid vector search strategy. The vector database and search are implemented using Qdrant. This hybrid approach combines sparse and dense embeddings for each document, allowing for sematic similarity searches on both vector types. Dense embeddings are generated using Voyage AI, a proprietary service, while sparse embeddings are created with SPLADE. Furthermore, the data corpus is categorized based on its source and intended use. Data for each use case is pre-processed and siloed into separate collections. Each silo is handled by a dedicated tool responsible for vector search and text post-processing tailored to the specific data source.

Systematic curation of relevant documents for the GSCP vector store

Vector databases rely on document embeddings for indexing and retrieval based on semantic similarity. However, semantically similar documents, even if thematically unrelated, can have close vector representations, leading to improper partitioning of the search space (Figure 2a). This results in clustering similar documents from different sources together, particularly in narrow domains like breast cancer. Additionally, naive chunking of long documents can fragment context.

Figure 2b demonstrates the volume imbalance across silos in the vector database, highlighting the low volume of crucial guidelines compared to clinical trials and PubMed data. For example, the guidelines related to breast cancer treatment from the National Comprehensive Cancer Network has total of 17.1

thousand tokens whereas all the clinical trial documents have 3.4 million tokens. The collated information from PubMed has a total of 460 thousand tokens. Despite being much smaller in volume, the guidelines and the generic information were used to synthesize a lot of responses that is further illustrated in Figure 4.

The corpus used in this work had documents that were thematically similar and hence, the vector embeddings from two separate but similar topics were naturally similar to each other. This created an issue as often guidelines from different sources would get intermixed as there was a significant overlap in the vector search space for these guidelines. To address the challenges associated with an overlapping search space and imbalanced data volume, we propose a Siloed Abstractive Vector Store (SVS) system. The SVS leverages collections within the vector database to partition the corpus into thematically distinct silos, such as NCCN Guidelines, ASCO Guidelines, PubMed, PharmGKB, Clinical Trials, and others. Each silo employs a dedicated retrieval module with tailored filtering, processing, and summarization techniques. The agent utilizes tools for asynchronous access to each silo, preventing information intermixing. This ensures, for example, that responses involving guidelines are synthesized based on information from a specific silo without contamination from other guidelines. This siloed approach allows the agent to present accurate information from various sources within a single coherent response.

Furthermore, each silo document includes a summarized abstraction alongside its chunked content. These summaries and chunks are linked via metadata. When a chunk is retrieved, its corresponding summary is retrieved simultaneously. These summaries are context-specific based on the silo's purpose. For instance, a summary recommending relevant clinical trials might differ from a summary highlighting a landmark drug trial, even if both originate from the same clinical trial document stored in separate silos. In our hand, this abstraction empowers the agent to grasp document semantics more effectively, reducing the number of reason-action loops required for response generation. For documents where chunk order is crucial, the summary guides the agent in re-ordering chunks using an LLM. Finally, depending on the utilized tool, retrieved document chunks undergo further LLM refinement before being forwarded to the agent, along with the document source metadata.

GSCP agent improves response quality over general purpose LLMs and RAGs

Agentic RAG systems act as a powerful approach for question answering tasks, particularly in the medical domain where access to comprehensive and informative answers is crucial. However, evaluating the effectiveness of these systems, especially in comparison to standalone LLMs or simpler RAG configurations, requires a rich and diverse set of evaluation datasets. Such datasets should encompass a variety of question formats, difficulty levels, and domains to provide a rigorous assessment of both context retrieval and response generation capabilities.

The absence of dedicated breast cancer question-answering datasets necessitated the creation of a comprehensive evaluation suite. We combined publicly available medical question-answering datasets with domain-specific samples encompassing both objective (multiple choice) questions from sources like MedMCQA[20] and MedQA[21], and subjective (open ended) questions from sources like PubMedQA[22] and internally constructed questions, all related to breast cancer (Figure 3a). For objective dataset, a Python script was utilized to identify all potential answer choices within the dataset for each question. Subsequently, we searched the corresponding responses for the presence of these choices. This approach was essential as the correct answer was frequently not explicitly listed among the provided options, thereby simplifying the downstream evaluation task.

We employed state-of-the-art LLM services, GPT-4 and Claude Opus-3, known for their comparable performance on benchmark tests (Figure 3b). The evaluation explored both simple RAG and agentic configurations.

For *Objective Question Answering* (QA) containing 223 questions, four quantitative metrics (accuracy, precision, recall, F1-score) were used to assess performance. Agentic systems significantly outperformed both RAG and standalone LLMs across all metrics. The GPT-4 powered agentic setup consistently achieved the highest scores in accuracy, recall, F1-score. Both LLM services demonstrated comparable performance, with GPT-4 exhibiting a slight edge. Agentic systems significantly outperformed both basic and standalone LLMs across all metrics. Agentic (GPT-4) consistently achieved the highest scores in accuracy, recall, F1-score, and precision (0.83, 0.83, 0.83, and 0.83 respectively), followed by the Agentic setup with Claude

Opus 3 with moderate success. Basic RAG models showed mixed results, with relatively strong precision but lower accuracy, recall, and F1-score. Standalone LLMs exhibited the lowest performance across all metrics.

For *Subjective QA* consisting of 113 questions, the DeepEval framework evaluated retrieval and generation performance for subjective questions. Standalone LLMs were excluded due to the absence of a retrieval context in their responses. Context precision, context relevancy, faithfulness, and answer relevancy metrics were employed. Agentic systems achieved superior performance in both retrieval and generation tasks. In terms of context precision, *Agentic (Claude Opus 3)* led with a score of 0.44, followed by *Agentic (GPT-4)* at 0.37, while basic models struggled. Both Agentic models achieved high context relevancy scores of 0.27, significantly surpassing the basic RAG models. For answer relevancy, *Agentic (GPT-4)* excelled with a score of 0.96, followed by *Agentic (Claude Opus 3)* at 0.90, while basic RAG models performed reasonably well. Finally, while faithfulness scores were comparable overall, *Agentic (Claude Opus 3)* achieved a perfect score of 1.

A custom in-house dataset focusing on precision oncology and breast cancer genetics, containing 25 questions, was created to simulate real-world healthcare complexities. This dataset, *Precision Oncology QA*, mimicked genetic markers, disease progression, and personalized treatment options. The same metrics used for subjective questions were applied. Agentic systems significantly outperformed basic systems in both context precision and relevancy. *Agentic (Claude Opus 3)* achieved a precision score of 0.51 and a relevancy score of 0.82, while *Agentic (GPT-4)* scored 0.52 for precision and 0.80 for relevancy. In contrast, basic RAG systems showed lower scores, with *Basic (Claude Opus 3)* at 0.20 for precision and 0.55 for relevancy, and *Basic (GPT-4)* at 0.15 for precision and 0.55 for relevancy. While answer relevancy and faithfulness scores were comparable across models, *Agentic (Claude Opus 3)* demonstrated slightly higher faithfulness with a score of 0.98 compared to *Basic (Claude Opus 3)* at 0.85. These evaluations demonstrate that moving from a RAG configuration to the proposed Agentic setup will improve performance no matter which LLM service is being used.

To construct a faithful representation of the clinical setting, synthetic patient case studies were generated through a collaborative effort involving practicing oncologists and LLM services. Oncologists contributed essential clinical insights, ensuring the case studies accurately reflected real-world medical complexities. These expert-provided details, devoid of specific patient data, served as the foundation for the LLMs to craft comprehensive case studies. To further enhance the authenticity of these synthetic cases, oncologists were asked to review the generated case studies to validate their alignment with the real clinical reports. The performance of agentic models, specifically Claude Opus 3 and GPT-4, was assessed using these three fabricated patient health records. In these experiments, the models were tasked with formulating suitable treatment plans based exclusively on the presented patient data. The detailed model outputs along with can be found in Supplementary Information Section 1.

Decoding the coordination among tools that improve GSCP response quality.

A key strength of the GSCP agent lies in its ability to leverage a suite of specialized tools for information retrieval and processing within distinct topic silos. By segmenting the corpus and employing dedicated tools per topic, the agent exposes its thought process during response generation. This allows the end-user to understand the weight given to various information sources and how they contribute to the final response. Furthermore, this targeted approach ensures retrieval and processing of the most relevant content for each query. Figure 4 exemplifies how much the agent employs each tool during different evaluation scenarios. Notably, NCCN guidelines are heavily relied upon for synthesizing responses to objective case study questions. Across all evaluations, PubMed consistently features as a high-usage tool, serving to supplement information gleaned from NCCN guidelines. Figure 5 illustrates the step-by-step sequence of tool utilization. In most instances, response generation occurs within 3 steps, with some reaching 5 steps. This sequential breakdown offers insights into the agent's reasoning process and its adept use of specialized tools within the siloed architecture.

Discussion

Large language models have demonstrated considerable potential across various domains, including healthcare and biomedical research. However, limitations in transparency and robust evaluation

methodologies have hindered their full clinical integration. GeneSilico CoPilot (GSCP) addresses these challenges by proposing an agent-based framework that leverages the inherent reasoning capabilities of LLMs to plan and execute tasks within the healthcare domain. This work focuses on the specific domain of breast cancer, showcasing the advantages of the GSCP framework over standalone LLMs and Retrieval-Augmented Generation (RAG) systems in both generic question answering and precision oncology tasks.

Evaluations conducted across public and private datasets demonstrate the superiority of the proposed agent-based framework compared to traditional RAG systems. In precision oncology question answering, the GSCP achieved an improvement of up to 15.29% in answer faithfulness compared to RAGs. Retrieval metrics also showed significant improvement, with the GSCP system achieving up to 200.83% and 47.27% better performance in context precision and context relevancy, respectively. These results highlight the clear advantage of the agent's reasoning and retrieval mechanisms over basic RAG approaches. Similar improvements were observed in the subjective question answering dataset, where the GSCP agent achieved up to 93.65% and 2600% improvement in context precision and context relevancy, respectively. The agent's retrieval mechanism facilitates a more robust reasoning process, and by incorporating these reasoning steps into the response generation, the GSCP system enhances the trustworthiness and transparency of its answers.

The GSCP departs from simple RAGs by employing pre-processed documents. These documents undergo summarization and are tagged with markdown annotations to facilitate hierarchical chunking by topic. This allows the agent to generate more coherent responses by leveraging a deeper understanding of the document structure and content. Furthermore, the GSCP system utilizes a suite of specialized tools, each optimized for retrieving different information volumes through a combination of dense and sparse embedding techniques. This multifaceted approach empowers the agent to perform more effective information retrieval, ultimately leading to a more comprehensive planning process for response generation.

Our evaluation revealed that Claude Opus 3 produced well-structured responses that resonated with oncology experts, despite achieving lower overall evaluation scores compared to OpenAI models. While

Opus 3 exhibited slower response generation times, often requiring up to two minutes to complete a response using provided tools, its outputs were characterized by superior readability. In terms of medical accuracy, Opus 3 offered more detailed explanations, including comprehensive drug and dosage information, while OpenAI models produced simpler responses.

Although both systems demonstrated comparable levels of medical accuracy, the significant disparity in human-perceived readability suggests an inability of the DeepEval evaluation framework to fully capture the nuanced aspects of response quality, particularly when considering human factors such as readability. This finding underscores the limitations of relying solely on automated metrics to assess model performance, particularly in complex domains such as medicine. While medical accuracy is undeniably crucial, it is essential to recognize that it is not the sole determinant of response quality. A comprehensive evaluation should consider additional factors, such as response clarity, coherence, and overall clinical utility, as perceived by human experts.

The GSCP system's transparent planning process, which can be visualized through tool usage, provides valuable insights into the relative importance of information sources. For example, our observations indicate a clear preference for NCCN guidelines over American Society of Clinical Oncology (ASCO) and European Society for Medical Oncology (ESMO) guidelines. It is noteworthy that NCCN guidelines underwent a meticulous manual paraphrasing process to convert them into plain text while preserving the information conveyed in the original flowcharts. In contrast, ASCO and ESMO guidelines primarily relied on LLM-based summarization. PubMed also emerged as a significant information source. While PubMed offers a wealth of open-access articles containing general knowledge, our focused initial retrieval process effectively transformed the PubMed collection into a more specialized corpus tailored to the domain of oncology. Analysis of tool usage statistics can be leveraged to inform future optimizations of the data sources, potentially leading to the deprecation, consolidation, or replacement of certain sources based on their effectiveness within the agent's framework.

Future endeavors include expanding our testing to encompass real-life patient cases and evaluating the GSCP system's capabilities in therapeutic decision support. This necessitates the development and

implementation of robust and reproducible evaluation metrics. Current frameworks like DeepEval, which rely on LLM services for LLM evaluation, are susceptible to inconsistencies. Therefore, there is a pressing need for more sophisticated evaluation methods specifically designed to assess the planning and reasoning capabilities of LLMs.

The GSCP system currently faces some limitations in terms of processing speed. The tool usage and frequent communication with the LLM service contribute to a processing delay, with complex cases requiring up to two minutes for response generation. Additionally, the vector store resides on a basic setup, resulting in slow retrieval times. Future improvements will focus on accelerating the vector store using quantization techniques.

Developing a patient-specific treatment regimen requires meticulous evaluation of various factors, including the patient's medical history, comorbidities, prior treatments, and potential drug toxicities. This necessitates a comprehensive review of the patient's medical records, encompassing laboratory results, imaging studies, and medication history. A thorough understanding of the patient's current health status and any coexisting conditions is also essential. Once this data is collected, the physician can begin exploring treatment options aligned with established clinical guidelines from NCCN, ASCO, and ESMO. These guidelines provide evidence-based recommendations informed by the latest research and clinical experience. However, it is equally important to consider the patient's individual needs and preferences, as well as their eligibility for ongoing clinical trials offering access to potentially groundbreaking therapies. This complex decision-making process necessitates the synthesis of information from diverse sources. The GSCP CoPilot system addresses this challenge by leveraging its knowledge base to recommend personalized treatment plans for each patient case. This streamlines the physician's workflow, facilitates informed decision-making, and ultimately contributes to enhanced patient care.

In conclusion, this work demonstrates the potential of developing domain-specific agent-based RAG systems. By focusing on a particular domain, such as oncology, the system can be optimized to effectively process and generate information within the context of a vast and complex data landscape.

FIGURES

Figure 1: Illustration of the working of the GeneSilico CoPilot

A) Schematic workflow depicting the entire pipeline of the agentic framework for precision oncology. The first step involves the collection and pre-processing of diverse medical data sources including literature, clinical trials, drug information and treatment guidelines – these serve as the tools. The second step involves the retrieval process where to efficiently extract relevant information given a query by employing appropriate tool selection, re-ranking, summarization and hybrid search. Further, the information retrieved is fed into a ReAct Agent that implements a cycle of reasoning, action and observation to synthesize the response to the query. The final step involves the generation of the response containing the medical insights and recommendations that caters to the use cases such as personalized therapy recommendations, clinical trial suggestions, genomic data analysis, and patient summaries; B) An oncologist provides a patient case study and prompts the GSCP to recommend a treatment plan. Upon receiving this query, GSCP engages in a structured Reason-Action-Observation process and synthesize the response based on the patient's specific clinical details

Figure 2: Representation of the data in the vector database

A) The 3D U-Map visualizes the vector search space, highlighting overlaps and intersections among various topics; B) The distribution of information across diverse topics within the vector database, illustrated through token counts, offers a comprehensive view of the content richness and topical breadth

Figure 3: Assessment of the performance of GSCP on question answering tasks. Distribution of the QA dataset and performance of GSCP on three different types of question answering tasks

Figure 4: Representation of tool usage during the three different types of question answering tasks (from left to right: *Objective QA*, *Subjective QA* and *Precision Oncology QA*)

Figure 5: Methodical use of tools in step-by-step response synthesis. The Sankey charts demonstrate how different tools were utilized across every step for response generation and the number of steps taken for generating a response for the given tasks

METHODS

Data sources

GSCP leverages a collection of manually curated data sources specific to breast cancer, compiled with the support of practicing oncologists. These sources include standard breast cancer guidelines from the National Comprehensive Cancer Network (NCCN), American Society of Clinical Oncology (ASCO), and European Society for Medical Oncology (ESMO).

Targeted Drug and Gene Information: To incorporate relevant drug and gene information, a curated list of 68 genes (including HRR and pharmacogenomics genes) and their targeted drugs was compiled. A customized GeneSilico gene panel for breast cancer therapy recommendations was designed, encompassing these 68 genes. The selection criteria for these genes included: genes associated with therapies (FDA-approved, Phase 3, and Phase 4 clinical trials); genes with high research significance and frequent alterations in databases like Human Somatic Mutation Database (HSMD) (digitalinsights.qiagen.com/hsmd/) and cBioPortal (www.cbioportal.org); genes associated with homologous recombination repair (HRR) mechanism; pharmacogenomic (PGx) genes relevant to breast cancer; normalized codon length of genes; and key genes present in other somatic panels such as MSK-IMPACT (www.mskcc.org/msk-impact), Foundation Medicine CDx diagnostic panel (https://www.foundationmedicine.in/our-services/cdx.html), and MedGenome panel (diagnostics.medgenome.com). The rankings from these criteria were combined using a rank aggregation algorithm to determine the final list of top genes, which were then manually validated. Pathogenic and likely pathogenic variants in breast cancer were selected using the HSMD, COSMIC[23], and ClinVar[24] databases. Additionally, the GeneSilico gene panel for breast cancer includes 32 microsatellite instability (MSI) hotspots.

Subsequently, this list was used to extract drug data from Drugbank Open Data (go.drugbank.com/releases/latest#open-data) [25], FDA drug labels (labels.fda.gov), RxList (www.rxlist.com), Therapeutic Target Database[26], Drugs.com (www.drugs.com), and Wikipedia. Web scraping was done using Selenium and Beautiful Soup. Drug approval details were obtained from the FDA and ClinicalTrials.gov. We

used the OpenFDA API and the Clinical Trials API to access the information. PubMed information was gathered using the PMC OA Web Service API. PharmGKB and JNCCN provided further breast cancer-specific data.

Data Abstraction and Summarization with Contextual Focus: To enhance context for lengthy documents, all documents were paraphrased using LLM services. Instead of generic summaries, task-specific summaries were created, extracting only relevant information. This context-aware process facilitated the summarization of pertinent sections rather than entire documents. The summarization was performed using Anthropic Claude Opus 3. The summarization prompt was provided depending on the requirement. For summarization of clinical trials for eligibility criteria, the prompt was "Make the following clinical trial information concise, highlighting the key eligibility criteria. Simply respond with the shortened text in markdown format." For clinical trials which contained drug approval information, the prompt was changed to "Give an abstract of the trial highlighting the drug approval information. Simply respond with the shortened text in markdown format." In every case, the summary was formatted with markdown tags. In our experiments, using markdown tags improved the response quality of the agent. Consequently, the retrieval module could fetch documents along with their summaries, improving retrieval performance by reducing the number of necessary documents and optimizing the context window. Although storing documents with multiple summaries creates redundancy across silos, this approach enhances agent performance. This summarization process was applied to all data sources containing long-form textual content.

Data Staging for Manageability and Retrieval: Data staging was implemented to improve manageability and retrieval efficiency. Before embedding, a copy of the data, along with extracted metadata and summaries, was stored in a NoSQL datastore. This was implemented using MongoDB. This simplifies and automates the embedding process while enabling retrievers to leverage full-text search on summaries for retrieving alternative results for a given query.

Manual Curation for Complex Documents: Certain documents, such as NCCN guidelines containing complex diagrams and flowcharts, underwent manual paraphrasing and conversion into plain text while

preserving the step-by-step narrative. Furthermore, NCCN documents were segmented based on cancer subtype, treatment phase, and treatment nature. Most other documents, including guidelines from ASCO or ESMO, were summarized using LLMs followed by manual inspection. Each summary was further segmented and annotated with markdown tags to enhance the agent's contextual understanding and facilitate the generation of more relevant responses.

Vector stores

A well-designed vector store is crucial for precise retrieval and minimizing the agent's reason-observation-action loops during response generation. Our implementation creates both dense and sparse embeddings for summaries and raw content. Dense embeddings, generated by a proprietary service – VoyageAI (https://www.voyageai.com/), capture nuanced data relationships. Sparse embeddings were created using SPLADE[27], enhance the search by focusing on key features. This approach enables direct hybrid search on the vector store.

Hybrid search leverages both sparse and dense vector embeddings within the vector database, facilitating full vector hybrid search on the indices. Sparse vectors offer computational efficiency and capture key document features, while dense vectors capture more nuanced data relationships. The hybrid search is configured to retrieve n elements using sparse vector search and m elements using dense vector search (n > m, typically n = 20, m = 10). The result list is generated using Reciprocal Ranked Fusion (RRF).

The ReAct agent.

By constructing a well-organized and semantically searchable corpus, we designed a ReAct agent specifically tuned for precision oncology. This involved identifying the agent's functionalities and developing corresponding tools. Each tool is a combination of query engines, response synthesis systems, re-rankers, and post-processors, categorized as follows:

- **Clinical Trial Information Retrieval:** This tool retrieves summarized clinical trial information from documents relevant to the patient's medical history and the prompt. Focusing on breast cancer, it searches over 900 trials (recruiting and past) for details like eligibility, duration, status, and location. The retrieval process involves a hybrid search on summaries followed by re-ranking for refinement.

Finally, the agent fetches raw documents linked to the selected summaries to infer the most appropriate trial and synthesize a detailed response.

- **Drug and Therapy Information Retrieval:** A combination of tools addresses drug information needs. For generic drug inquiries, a single tool retrieves data on standard dosages, toxicities, and mechanisms of action from FDA drug labels, RxList, and other sources. Personalized responses leverage a combination of tools that consult various cancer treatment guidelines before response synthesis.
- **Precision Oncology Guideline Retrieval:** Given patient history context, a routing tool selects tools that fetch relevant NCCN, ASCO, and ESMO guidelines. These tools operate independently to provide individual patient-specific guidelines. Additionally, tools for generic literature search are invoked to supplement retrieved information.
- **Generic Literature Search:** This category encompasses manually curated sources like PubMed searches on drug-gene pairs and breast cancer therapies. Information is also retrieved from JNCCN and PharmGKB, providing comprehensive data on standard care protocols, drug information, patient outcomes, and generic therapy guidelines.

A complete configuration of these tools is presented as a JSON file in Supplementary Information Section 2.

Experimental setup

We evaluated the proposed method using datasets constructed from public sources and real-life cases. Standard public datasets for breast cancer are unavailable. Therefore, we extracted breast cancer-related questions from multiple sources and categorized them as subjective (requiring long-form answers) or objective (multiple-choice). A simple keyword-based search facilitated extraction, followed by manual review by practicing oncologists to ensure question correctness. The objective dataset comprised 223 questions from MedMCQA and MedQA (USMLE), while the subjective dataset consisted of 113 questions extracted from PubMedQA and an in-house dataset (InternalQA). For objective questions lacking a single clear answer where oncologists identified multiple correct options, the questions were reworked as

subjective ones. To simulate the complexities encountered by medical professionals in real-world oncology practice, we constructed a custom in-house dataset, the Precision Oncology dataset, of 25 questions focused on precision oncology and breast cancer genetics. This dataset embodies case-study like scenarios, mimicking an oncologist's investigative process. The internal dataset as well as the precision oncology datasets were created with the support of practicing oncologists. The dataset for evaluation is provided in Supplementary Data.

Accuracy, F1 score, precision, and recall were used to assess system performance for simple multiple-choice questions. For the subjective and precision oncology datasets, the DeepEval framework (https://docs.confident-ai.com/) evaluated our system and compared its performance to a RAG system. This framework employs Contextual Precision (ranking relevant information), Contextual Relevancy (overall retrieved context relevance), Faithfulness (factual alignment between response and retrieved context), and Answer Relevancy (ratio of relevant statements in the answer) to measure the retrieval and generation performance.

Data Availability

The datasets used in this study are available in the supplementary materials. Supplementary Data file contains evaluation datasets, including the objective questions extracted from MedMCQA and MedQA (USMLE), subjective questions from PubMedQA and our in-house dataset (InternalQA), and custom Precision Oncology dataset. Our in-house datasets (InternalQA and Precision Oncology) were created with the support of practicing oncologists and are included in the supplementary materials. The DeepEval framework used for performance evaluation is publicly accessible at https://docs.confident-ai.com/. Any additional data that supports the findings of this study are available from the corresponding author upon reasonable request.

Supplementary Files

Supplementary Information Section 1: Contains selected synthetic case studies for intervention plan. These case studies were designed by oncologists based on real-life cases. The file contains the entire

response provided by GSCP when presented with the case studies. The following system prompt was used: "Based on the provided case study, suggest a comprehensive treatment plan with evidence and cancer management approach for the patient, taking into consideration their condition, history, family history, and comorbidities. Outline the recommended course of action, available treatment options with evidence and follow up plan. Also, suggest if rebiopsy, genomics (NGS), additional biomarkers required or not. Additionally, identify relevant ongoing clinical trials that the patient may be eligible for, and provide survival statistics.

Break down the response into section and try to answer these questions:

What should be the next line of treatment?

Is genomics required in this case?

How should be the follow up?

Genetic counselling required?

Any role of immunotherapy?"

Supplementary Information Section 2: Contains the configuration file of the different tools used by the agent. The "top_k" and the "sparse_k" determine the number of results fetched based on the dense and the sparse vector match respectively. The "output_type" specifies how the LLM service in each tool should aggregate the information from the documents after the reranking process.

Supplementary Data: Consists of the questions and the corresponding ground truth for the datasets – Subjective, Objective and Precision Oncology.

REFERENCES

1. Haslam, A., Kim, M. S. & Prasad, V. Updated estimates of eligibility for and response to genome-targeted oncology drugs among US cancer patients, 2006-2020. *Annals of Oncology* 32, 926–932 (2021).

2. Thirunavukarasu, A. J. *et al.* Large language models in medicine. *Nature Medicine 2023 29:8* 29, 1930–1940 (2023).

3. Gilbert, S., Harvey, H., Melvin, T., Vollebregt, E. & Wicks, P. Large language model AI chatbots require approval as medical devices. *Nature Medicine 2023 29:10* 29, 2396–2398 (2023).

4. Zhao, H. *et al.* Explainability for Large Language Models: A Survey. *ACM Trans Intell Syst Technol* 15, 38 (2024).

5. Haltaufderheide, J. & Ranisch, R. The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). *npj Digital Medicine 2024 7:1* 7, 1–11 (2024).

6. Lee, P., Bubeck, S. & Petro, J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *N. Engl. J. Med.* 388, 1233–1239 (2023).

7. Singhal, K. *et al.* Large language models encode clinical knowledge. *Nature 2023 620:7972* 620, 172–180 (2023).

8. Luo, R. *et al.* BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief Bioinform* 23, 1–11 (2022).

9. Lee, J. *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 1234–1240 (2020).

10. Yang, X. *et al.* A large language model for electronic health records. *npj Digital Medicine 2022 5:1* 5, 1–9 (2022).

11. Sorin, V. *et al.* Large language model (ChatGPT) as a support tool for breast tumor board. *npj Breast Cancer 2023 9:1* 9, 1–4 (2023).

12. Yalamanchili, A. *et al.* Quality of Large Language Model Responses to Radiation Oncology Patient Care Questions. *JAMA Netw Open* 7, e244630–e244630 (2024).

13. Rydzewski, N. R. *et al.* Comparative Evaluation of LLMs in Clinical Oncology. *NEJM AI* 1, (2024).

14. Lazris, D., Schenker, Y. & Thomas, T. Exploring AI-generated content and professional guidelines in cancer symptom management: A comparative analysis between ChatGPT and NCCN guidelines. *Journal of Clinical Oncology* 42, e13610–e13610 (2024).

15. Zhou, S., Wang, N., Wang, L., Liu, H. & Zhang, R. CancerBERT: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. *Journal of the American Medical Informatics Association* 29, 1208–1216 (2022).

16. Tan, R. *et al.* Retrieval-augmented large language models for clinical trial screening. *https://doi.org/10.1200/JCO.2024.42.16_suppl.e13611* 42, e13611–e13611 (2024).

17. Ferber, D. *et al.* GPT-4 for Information Retrieval and Comparison of Medical Oncology Guidelines. *NEJM AI* 1, (2024).

18. Li, Y. *et al.* RefAI: a GPT-powered retrieval-augmented generative tool for biomedical literature recommendation and summarization. *Journal of the American Medical Informatics Association* (2024) doi:10.1093/JAMIA/OCAE129.

19. Yao, S. *et al.* ReAct: Synergizing Reasoning and Acting in Language Models. (2022).

20. Pal, A., Umapathi, L. K. & Sankarasubbu, M. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. *Proceedings of Machine Learning Research* vol. 174 248–260 Preprint at https://proceedings.mlr.press/v174/pal22a.html (2022).

21. Jin, D. *et al.* What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. *Applied Sciences 2021, Vol. 11, Page 6421* 11, 6421 (2021).

22. Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W. & Lu, X. PubMedQA: A Dataset for Biomedical Research Question Answering. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural*

Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference 2567–2577 (2019) doi:10.18653/v1/d19-1259.

23. Bamford, S. *et al.* The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *British Journal of Cancer 2004 91:2* 91, 355–358 (2004).

24. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* 42, D980–D985 (2014).

25. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 46, D1074–D1082 (2018).

26. Chen, X., Ji, Z. L. & Chen, Y. Z. TTD: Therapeutic Target Database. *Nucleic Acids Res* 30, 412–415 (2002).

27. Formal, T., Piwowarski, B. & Clinchant, S. SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking. *SIGIR 2021 - Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* 2288–2292 (2021) doi:10.1145/3404835.3463098.

ACKNOWLEDGEMENTS

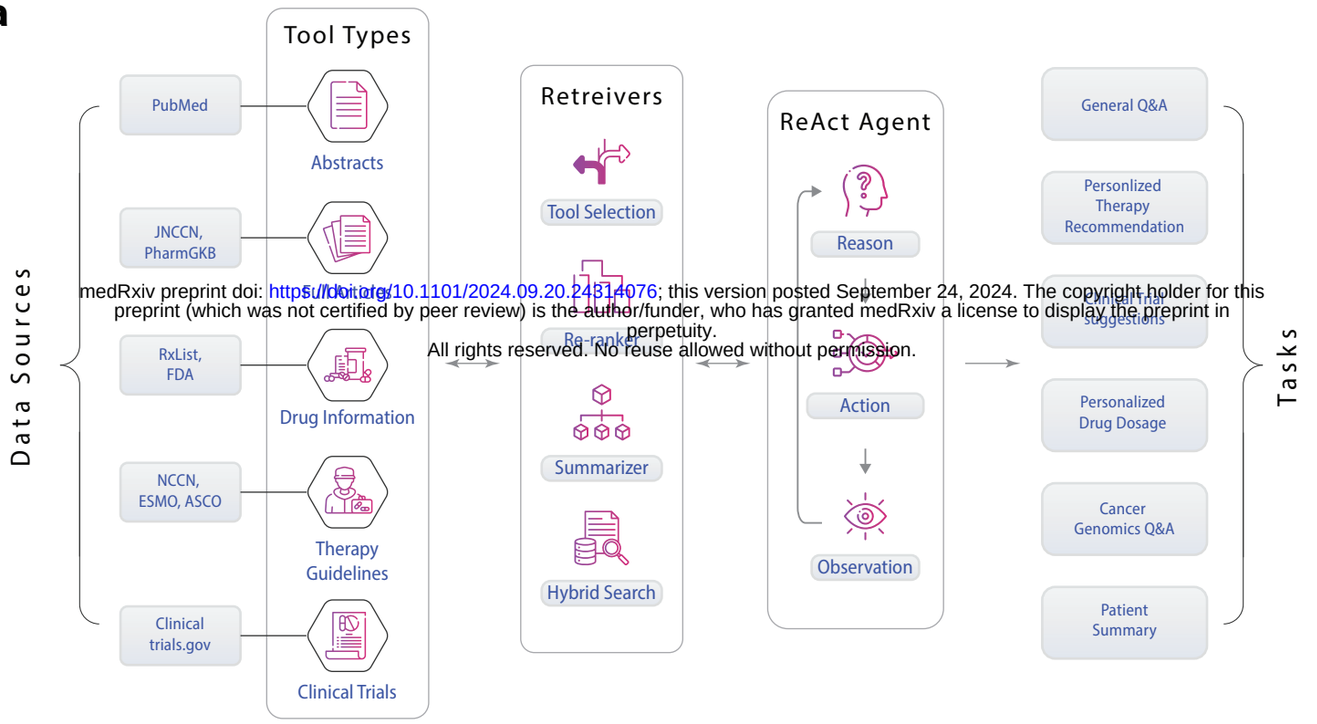
The authors would like to thank Claire Cohen from GeneSilico Inc., and Ruhani Bhatia for their valuable input.

AUTHOR CONTRIBUTION STATEMENT

DS conceived the study. RD and KM designed and implemented the agent and performed all experiments under the supervision of DS. SS, JN, and SM defined the scope for genetic data ingestion. JN also performed evaluations and generated diagrams. NA implemented the vector database and MongoDB, contributed to the agent design, and led the productionization efforts. AP assisted with vector database implementation, conducted data scraping and collation, and developed tools for agent use. VU and AS developed and deployed the copilot web application in production, with support from TS and HC. RT, SP, and SN provided clinical input to ensure the relevance of data and outputs. RT further contributed to the design of clinical questions and responses. UM, SB, and MF provided critical feedback on the agent design. All authors reviewed and approved the final manuscript.

Figure 1

a



b

Based on the provided case study, suggest a comprehensive treatment plan with evidence and cancer management approach for the patient, taking into consideration their condition, history, family history, and comorbidities. Outline the recommended course of action, available treatment options with evidence and follow up plan. Also, suggest if rebiopsy, genomics (NGS), additional biomarkers required or not. Additionally, identify relevant ongoing clinical trials that the patient may be eligible for, and provide survival statistics.

Patient case study

- Mrs. X, a 48-year-old premenopausal female, diagnosed as carcinoma right breast cT3N1, Er 8+ Pr 7+ Her2 3+diagnosed in July 2023. Staging Pet-CT showed no metastasis to other organs. Family history of ovary cancer present in maternal aunt. No comorbidities.
- She received 6 cycles neoadjuvant chemotherapy with TCHP regimen - Docetaxel, carboplatin, Transtuzumab, Pertuzumab) and underwent Breast conservation surgery in February 2024.
- Histopathology revealed residual Invasive Ductal Carcinoma of Right Breast with free margins and 2/16 axillary lymph nodes positive for tumour. Pathologic stage (8th AJCC Staging system): ypT1cN1a
- She completed adjuvant Radiation, and Transtuzumab, Pertuzumab regimen x 6 doses. She was put on ovary suppression regimen with Inj Goserelin monthly and started on adjuvant Letrozole along with calcium and vitamin D supplements.
- Presently she is admitted for seventh adjuvant Transtuzumab and Pertuzumab.

Reason - Action - Observation Process

Source: NCCN
Observation: Based on the information provided, the recommended treatment for this 48-year-old premenopausal female with stage cT3N1 ER+/PR+/HER2+ breast cancer who received neoadjuvant TCHP (docetaxel, carboplatin, trastuzumab, pertuzumab) chemotherapy is:
- Since she had residual disease after neoadjuvant therapy with 2/16 positive nodes at surgery, ado-trastuzumab emtansine (T-DM1) alone is recommended ...

Source: ASCO
Observation: Based on the provided context, there are no specific recommendations for rebiopsy, genomic testing like next-generation sequencing (NGS), or additional biomarkers for a premenopausal patient with residual ER+/PR+/HER2+ breast cancer after neoadjuvant chemotherapy.

Source: Clinical Trials
Observation: Based on the provided clinical trial information, a few trials may be relevant for a premenopausal patient with residual ER+/PR+/HER2+ invasive breast cancer after neoadjuvant chemotherapy and HER2-targeted therapy:
1. NCT05720260 - This phase 2 trial is investigating different treatment combinations, including hormone therapy, AKT inhibitor, and immunotherapy, in premenopausal or perimenopausal women with ER+ locally advanced or metastatic breast cancer. Key inclusion criteria include premenopausal status and ER positivity >1%.
2. NCT05800756 - This trial includes premenopausal women aged ≥18 years with pathologically confirmed ER+/PR+/HER2+ invasive breast cancer. ....

Source: Clinical Trials Publications
Observation: The most relevant information is from the KATHERINE trial, which found that among patients with HER2+ early breast cancer who had residual invasive disease after neoadjuvant therapy, the estimated percentage free of invasive disease at 3 years was 88.3% in the T-DM1 group compared to 77.0% in the trastuzumab group. The risk of recurrence or death was 50% lower with adjuvant T-DM1 compared to trastuzumab alone.

Response

Based on the provided case details and the information gathered from guidelines and clinical trials, the recommended treatment plan for Mrs. X, a 48-year-old premenopausal female with ER+/PR+/HER2+ breast cancer who had residual disease after neoadjuvant TCHP chemotherapy, is as follows:

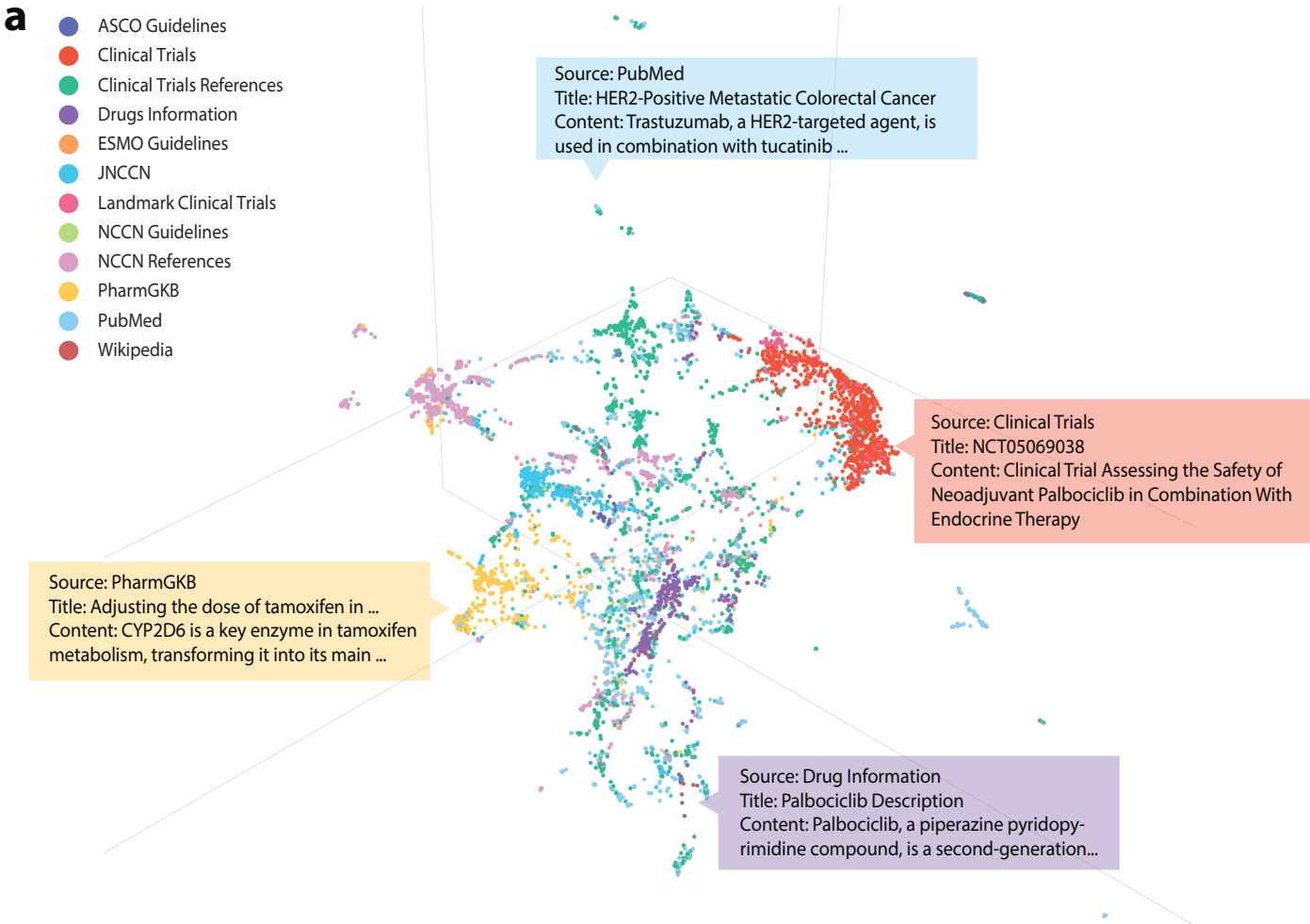
1. Complete a total of one year of HER2-targeted therapy by switching to adjuvant T-DM1 for 14 cycles, since she had residual disease and positive nodes after neoadjuvant treatment. This is supported by the KATHERINE trial which showed a 50% reduction in recurrence or death with T-DM1 compared to trastuzumab alone in patients with residual invasive disease.
2. Continue ovarian suppression with monthly goserelin injections and adjuvant letrozole for at least 5 years, as she is premenopausal with ER+/PR+ disease. Ensure she takes calcium and vitamin D supplements to maintain bone health while on endocrine therapy.
3. Monitor cardiac function with periodic echocardiograms or MUGA scans while receiving HER2-targeted therapy, as both trastuzumab and T-DM1 can affect left ventricular ejection fraction.
4. Follow up with regular oncology visits every 3-6 months for the first 2-3 years, then every 6-12 months for years 4-5 to monitor for recurrence. Annual mammograms should be performed to screen the contralateral breast.

At this time, there is no clear indication for rebiopsy, additional genomic testing, or checking further biomarkers, as the current standard of care is to proceed with adjuvant T-DM1 based on the finding of residual disease after neoadjuvant treatment.

Mrs. X may be eligible for some clinical trials investigating novel therapies in HER2+ breast cancer, such as NCT05720260 and NCT05800756, although her stage is not as advanced as specified in some trial criteria. While exact survival statistics are not available for patients matching her characteristics, data from KATHERINE suggests that 88.3% of patients receiving adjuvant T-DM1 remain free of invasive disease at 3 years.

In summary, the key elements of Mrs. X's treatment plan are adjuvant T-DM1, ongoing ovarian suppression with letrozole, regular cardiac monitoring, and close follow-up for recurrence surveillance. This evidence-based approach offers the best chance of long-term disease control based on current guidelines and available clinical trial data.

Figure 2


a
ASCO Guidelines
Clinical Trials
Clinical Trials References
Drugs Information
ESMO Guidelines
JNCCN
Landmark Clinical Trials
NCCN Guidelines
NCCN References
PharmGKB
PubMed
Wikipedia
Source: PubMed
Title: HER2-Positive Metastatic Colorectal Cancer
Content: Trastuzumab, a HER2-targeted agent, is used in combination with tucatinib ...
Source: Clinical Trials
Title: NCT05069038
Content: Clinical Trial Assessing the Safety of Neoadjuvant Palbociclib in Combination With Endocrine Therapy
Source: PharmGKB
Title: Adjusting the dose of tamoxifen in ...
Content: CYP2D6 is a key enzyme in tamoxifen metabolism, transforming it into its main ...
Source: Drug Information
Title: Palbociclib Description
Content: Palbociclib, a piperazine pyridopy-rimidine compound, is a second-generation...

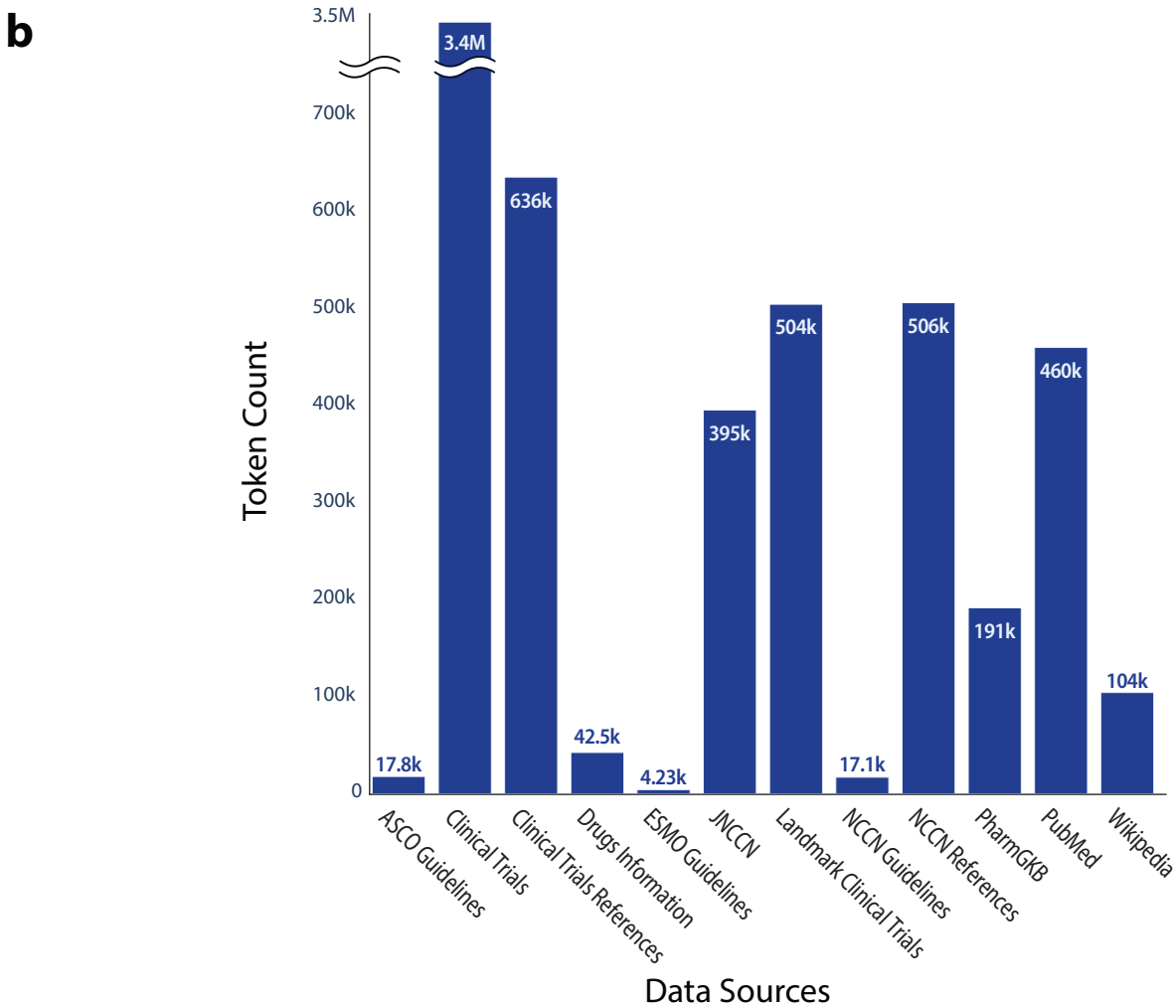

b
Token Count
3.5M
700k
600k
500k
400k
300k
200k
100k
0
17.8k
3.4M
636k
42.5k
4.23k
395k
504k
17.1k
506k
191k
460k
104k
ASCO Guidelines
Clinical Trials
Clinical Trials References
Drugs Information
ESMO Guidelines
JNCCN
Landmark Clinical Trials
NCCN Guidelines
NCCN References
PharmGKB
PubMed
Wikipedia
Data Sources

Figure 3

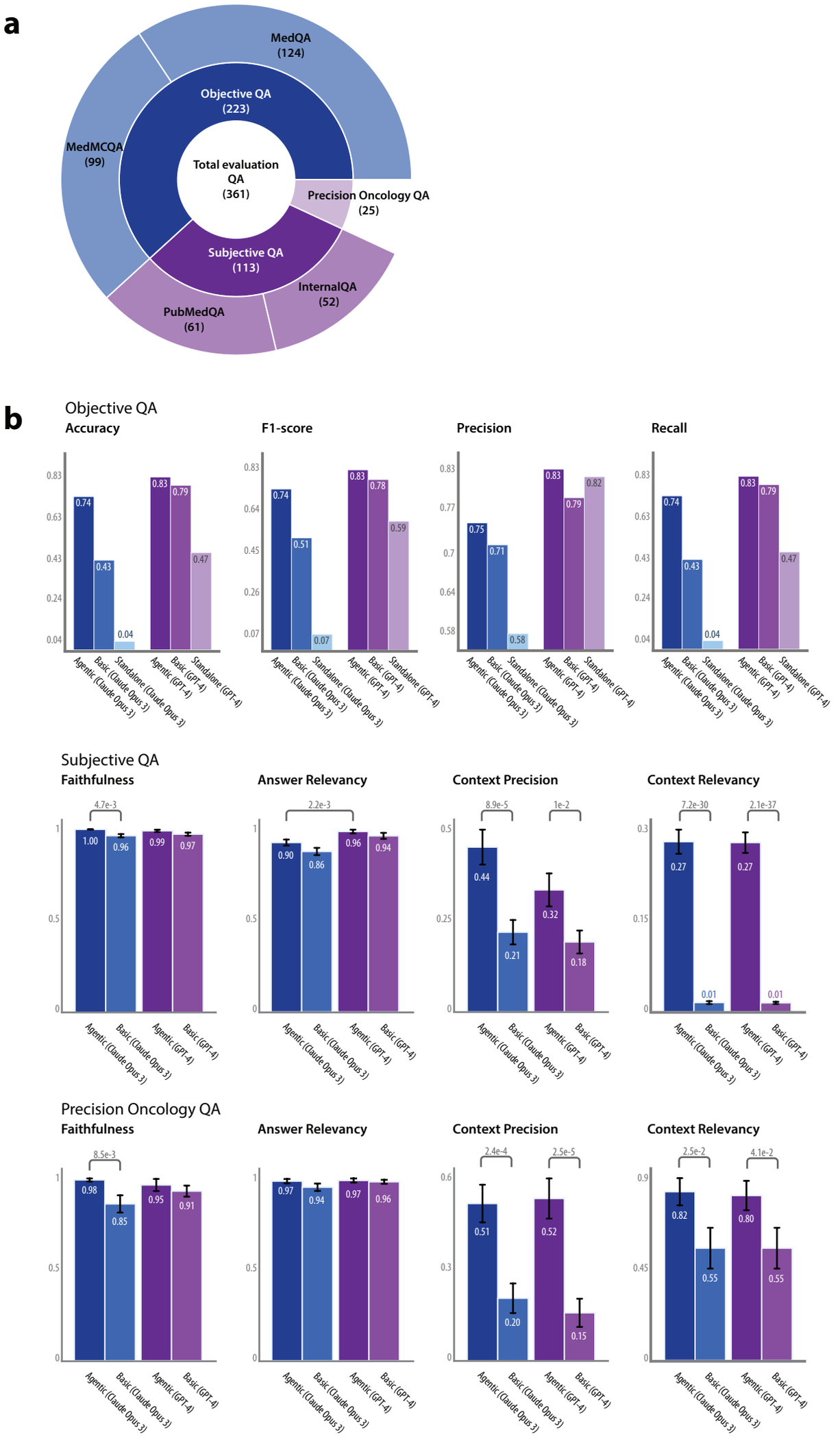


Figure 4

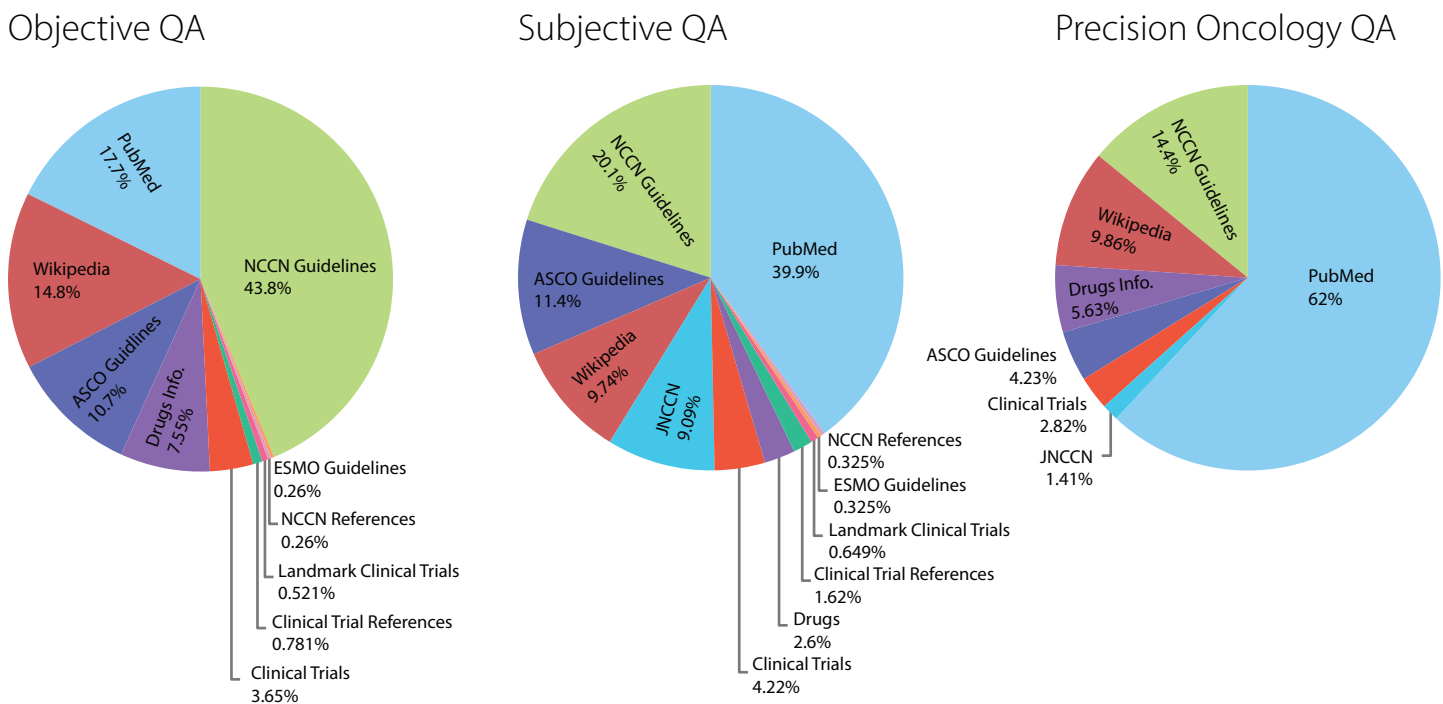

Objective QA
PubMed 17.7%
NCCN Guidelines 43.8%
Wikipedia 14.8%
ASCO Guidlines 10.7%
Drugs Info. 7.55%
ESMO Guidelines 0.26%
NCCN References 0.26%
Landmark Clinical Trials 0.521%
Clinical Trial References 0.781%
Clinical Trials 3.65%
Subjective QA
NCCN Guidelines 20.1%
PubMed 39.9%
ASCO Guidelines 11.4%
Wikipedia 9.74%
JNCCN 9.09%
NCCN References 0.325%
ESMO Guidelines 0.325%
Landmark Clinical Trials 0.649%
Clinical Trial References 1.62%
Drugs 2.6%
Clinical Trials 4.22%
Precision Oncology QA
NCCN Guidelines 14.4%
Wikipedia 9.86%
Drugs Info. 5.63%
PubMed 62%
ASCO Guidelines 4.23%
Clinical Trials 2.82%
JNCCN 1.41%

Figure 5