

Design and Development of an NLP Pipeline for Therapeutic Education in Oncology

A Project Report

submitted in partial fulfillment of the requirements for the

End of year project

in

Industrial IT and Automation Engineering (IIA 4)

by

Malek Zitouni

Mohammed Karim Ben Boubaker

Jaouadi Mahdi

Under the supervision of

Mrs. BOUBAKER Olfa



Department of Physics and Instrumentation Engineering

National Institute of Applied Science and Technology

Tunis, Tunisia.

June 2025

Acknowledgement

We appreciate our supervisor, Mrs. BOUBAKER Olfa, for her great patience and valuable insights. We are also thankful to everyone who provided advice and encouragement throughout this project.

Additionally, we want to express our gratitude to our university for giving us the necessary resources and environment to be able to complete this project. We have greatly benefited from the guidance and knowledge shared by our instructors and peers, who have been essential in guiding us on our journey.

Contents

1 Introduction.....	3
1.1 Problem Statement.....	3
1.2 Technological Motivation.....	4
1.3 Research Objectives.....	4
1.4 Technical Specifications.....	4
2 System Design and Implementation.....	5
2.1 RAG Architecture Foundation and Patent Analysis.....	5
2.1.1 Innovation Framework and Problem Resolution.....	5
2.1.2 Encoder Architecture and Semantic Representation.....	5
2.1.3 Retrieval Optimization and Computational Efficiency.....	7
2.1.4 Augmented Generation and Context Management.....	7
2.2 Architecture of the NLP Pipeline.....	8
2.2.1 System Overview and Design Philosophy.....	8
2.3 Semantic Relevance Validation Layer.....	9
2.4 Data Collection process.....	10
2.4.1 Sources Selection Process.....	10
2.4.2 Methods and Techniques for Transforming Data.....	10
2.4.3 Q&A File Generation.....	11
3 Results and Conclusion.....	13
3.1 Evaluation metrics definition.....	13
3.2 Evaluation Process.....	13
3.2 Results and Interpretation.....	14
4 References.....	16
5 Data Sources.....	17
6 List of Figures.....	18
7 List of Tables.....	18
8 Appendix.....	19

1 Introduction

1.1 Problem Statement

The oncology healthcare sector faces a critical challenge characterized by escalating provider burnout and inefficient resource allocation. Recent longitudinal studies demonstrate a concerning trend in oncology provider wellness, with burnout rates among oncology physician assistants increasing from 34.8% in 2015 to 48.7% by 2019 ([Jennifer Byrne et al., 2020](#)). This deterioration in provider well-being directly correlates with reduced quality of patient care and increased healthcare system costs.

The Association Tunisienne d'Oncologie has identified a significant operational bottleneck where oncologists routinely encounter repetitive patient inquiries that consume substantial consultation time. These queries, while seemingly straightforward, often require nuanced medical assessment based on patient-specific factors, treatment regimens, and associated risks. For instance, questions such as "Can a cancer patient take a hot shower?" demand careful assessment of treatment-induced immunosuppression, skin integrity, medication interactions, and individual risk factors.

Across healthcare, doctors often spend over half their time on data entry and admin tasks instead of making complex clinical decisions ([McKesson et al., n.d.](#)), making this problem even more difficult. This misuse of specialist expertise represents a fundamental system design flaw that compromises both provider satisfaction and patient outcomes.

By offloading repetitive queries to a clinically grounded RAG system, oncologists can reclaim valuable consultation time, reduce burnout, and focus on complex, high-impact care.

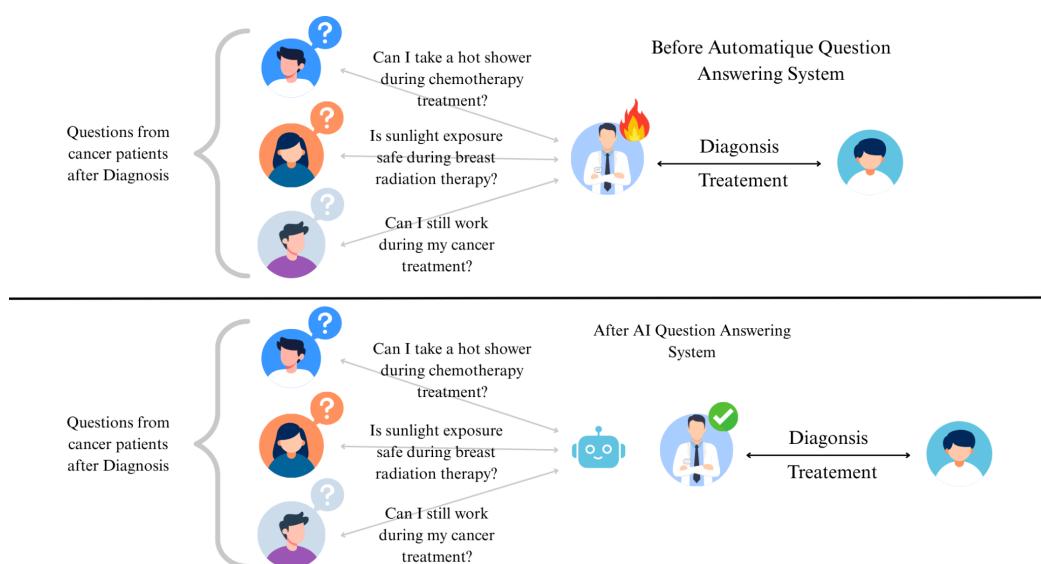


Figure 1 - Current Clinical Workflow vs. Proposed RAG-Enhanced

1.2 Technological Motivation

The rise of large language models (LLMs) and retrieval-augmented generation (RAG) systems presents novel opportunities for healthcare process automation. RAG systems mitigate the limitations of standalone LLMs by grounding responses in external, clinically validated knowledge sources ([Amugongo et al., 2024b](#)). This approach proves particularly relevant for healthcare applications where evidence-based information delivery is essential for patient safety and clinical efficacy.

RAG systems enhance generative AI accuracy by employing dynamic information retrieval from curated, domain-specific data sources. This capability enables the development of healthcare information systems that can provide medically accurate responses while maintaining complete traceability to trusted clinical sources. The integration of external knowledge bases ensures that generated content reflects current medical standards and established clinical guidelines.

Recent implementations of RAG-based healthcare education systems demonstrate practical viability for delivering accessible, culturally relevant medical information ([Al Ghadban et al., 2023](#)).

1.3 Research Objectives

The main objective of this research is to develop and validate a RAG-based natural language processing system that can use physician-curated knowledge sources to generate evidence-based answers to questions from cancer patients. This system will serve as an intelligent intermediary between patients and healthcare providers, filtering routine inquiries while ensuring medically appropriate guidance.

Secondary objectives focus on critical technical and clinical requirements. The system must recognize the limits of its medical knowledge and clearly reject questions it cannot answer safely, with clear rejection procedures for out-of-scope queries.

1.4 Technical Specifications

The proposed system requires implementation of several functional capabilities to meet clinical deployment standards. Core question-answering functionality must utilize physician-provided or trusted internet sources to ensure medical accuracy. Personalized response generation based on patient demographics, medical history, and current treatment status will enable individualized guidance delivery. Knowledge boundary detection with explicit refusal mechanisms prevents inappropriate medical advice for queries outside the system's validated knowledge domain.

2 System Design and Implementation

2.1 RAG Architecture Foundation and Patent Analysis

2.1.1 Innovation Framework and Problem Resolution

Microsoft's Retrieval-Augmented Generation framework, outlined in Microsoft's patent US20240346256A1 ([Qin, 2024](#)), effectively solves three key problems that limit traditional large language models. Knowledge staleness represents the primary challenge where traditional LLMs rely on static training data with fixed cutoff dates ([Microsoft, 2025](#)). RAG resolves this limitation by integrating dynamic information retrieval systems that access current knowledge without requiring model retraining.

Traditional LLMs face a major challenge with computational costs, as updating their knowledge requires expensive retraining of the entire model. RAG addresses this by implementing an efficient information retrieval system that integrates new knowledge with existing training data, eliminating the need for costly retraining while preserving performance. Furthermore, RAG improves output quality by anchoring its responses in verified data sources, which substantially reduces the generation of false information while maintaining natural, coherent language ([Amazon Web Services, n.d.](#)).

200

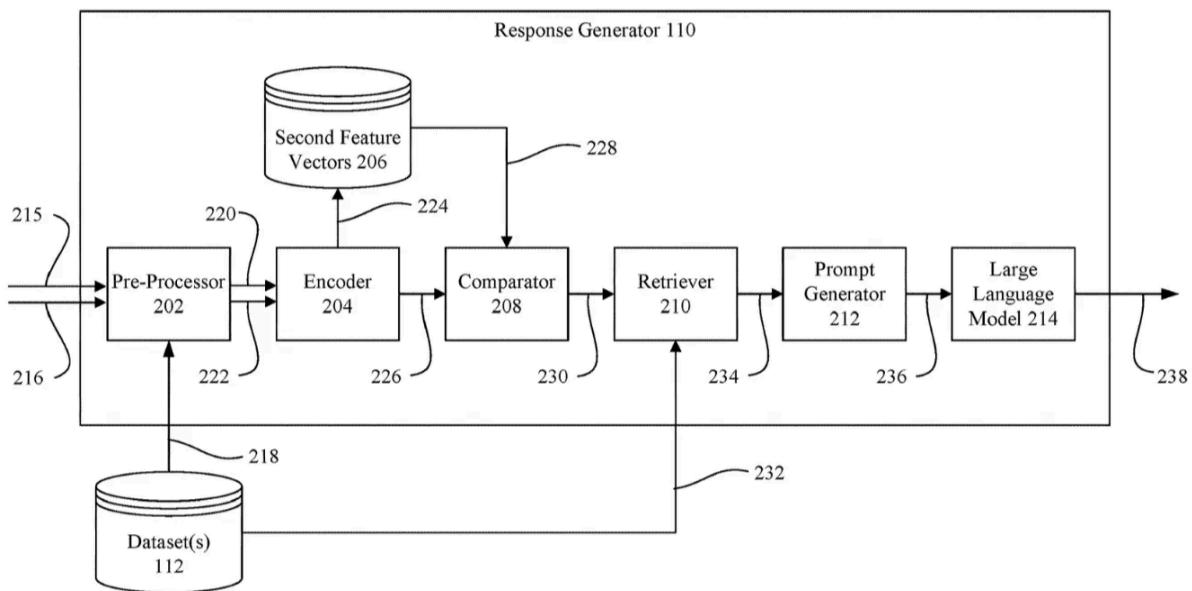


Figure 2 - Retrieval Augmented Generation (RAG) Architecture (per Microsoft Patent US20240346256A1)

2.1.2 Encoder Architecture and Semantic Representation

Encoders serve as the semantic bridge between natural language queries and retrievable knowledge fragments, transforming textual input into dense vector

representations within high-dimensional semantic spaces. Given an input text sequence T , the encoder produces a dense vector embedding:

$$E(T) = f(T) \in R^d$$

Where f represents the encoding function and d denotes the embedding dimension (typically 384, 768, or 1536). This transformation enables mathematical similarity computations between user queries and knowledge base documents through vector operations such as cosine similarity. ([MyScale, 2024](#))

$$\text{sim}(q, d) = (E(q) \cdot E(d)) / (\|E(q)\| \cdot \|E(d)\|)$$

Where q denotes the query and d denotes a document in the knowledge base.

BERT (Bidirectional Encoder Representations from Transformers) revolutionized natural language processing by pre-training deep neural networks on vast amounts of text data. Unlike previous models, BERT processes words in relation to all other words in a sentence, capturing **bidirectional context**. This approach allows it to generate context-aware **word embeddings**, which are crucial for understanding the nuances of human language ([Devlin et al., 2018](#)).

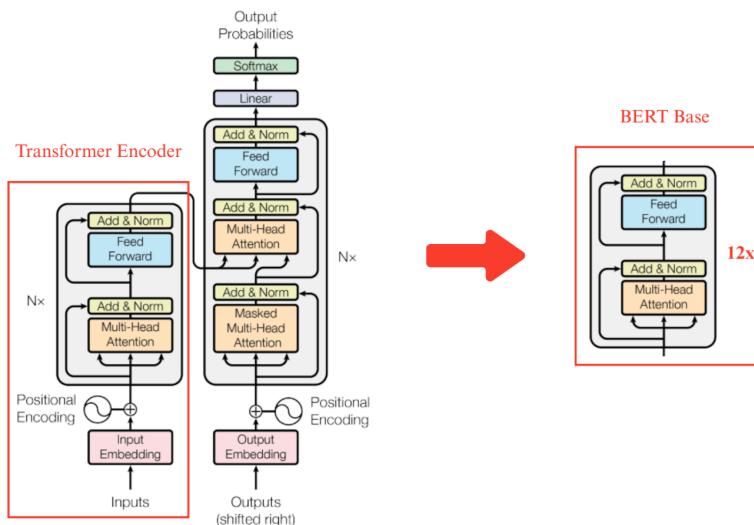


Figure 3 - BERT Architecture Inspired by the Transformer Encoder presented in Attention is all you need paper ([Vaswani et al., 2017](#))

Sentence-BERT (SBERT) represents a significant advancement in open-source encoding solutions, specifically trained to maximize similarity between pooled sentence representations rather than individual tokens. This approach addresses the fundamental limitation of basic BERT architectures, which generate embeddings for individual tokens rather than complete semantic units. SBERT's sentence-level semantic representations prove essential for document retrieval applications where meaning must be captured at the sentence and paragraph level ([Efimov, 2023](#)).

2.1.3 Retrieval Optimization and Computational Efficiency

Simple retrieval methods Check how similar a query is to *every single item* in the knowledge base. This "exhaustive" comparison means the time it takes grows directly with the number of items (complexity). For very large knowledge bases, this becomes too slow and impractical.

$$Cost = O(n \cdot d)$$

where d represents the embedding dimension. This linear scaling presents fundamental scalability barriers for practical deployment scenarios ([Aumüller et al., 2018](#)).

To solve the issue of slow searches, **Approximate Nearest Neighbor (ANN)** algorithms come into play. These methods speed up the search process significantly, bringing the computational cost down to roughly $O(\log n)$. They achieve this by using clever ways to organize the data, like creating a hierarchy or an index. The search then focuses on finding the 'k' documents that are most similar to the query (["Nearest neighbor search," n.d.](#)).

Vector databases like Pinecone, Weaviate, and Chroma are key tools for this. They're designed to efficiently store and retrieve these complex 'embedding' numbers. These systems use smart indexing and distributed computing to handle many searches at once, which is essential for real-world healthcare uses.

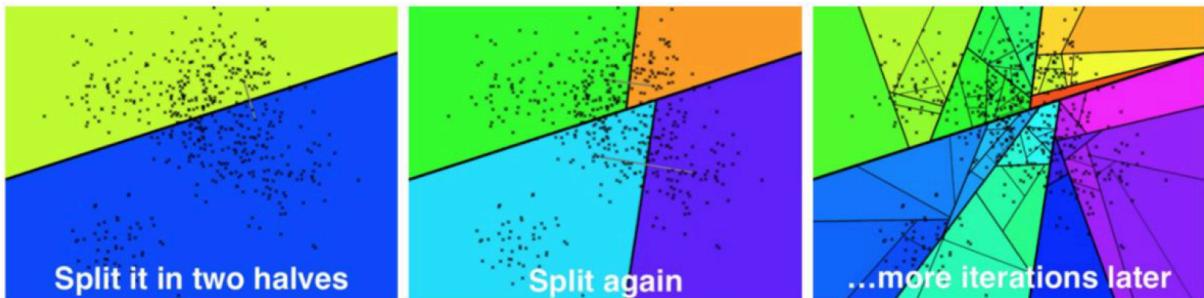


Figure 3 - Process of splitting in ANN similar to how binary search works ([SDS-AAU, n.d.](#))

2.1.4 Augmented Generation and Context Management

Retrieval Augmented Generation (RAG) systems strategically use the transformer architecture's context window as a dynamic, short-term memory mechanism. This is achieved by programmatically injecting retrieved, semantically relevant documents as contextual information into the LLM's input, alongside the user's initial query. The construction of this augmented prompt (Paugmented) is formalized as a concatenation:

$$P_{augmented} = [C_1, C_2, C_3, \dots, C_k, Q]$$

Where C_i denotes an individual retrieved context document (or passage), and Q represents the user's original query. This augmented input then drives the generative process, which can be expressed as

$$Response = LLM(P_{augmented}) = LLM([Contexts, Query])$$

This fundamental shift transforms the LLM's response generation from relying purely on its **parametric knowledge** $P(\text{response} \mid \text{query})$ to a **hybrid parametric-retrieval synthesis** $P(\text{response} \mid \text{query}, \text{retrieved contexts})$. This enables the LLM to access and integrate information *beyond* its original training dataset, thereby enhancing factual accuracy and currency.

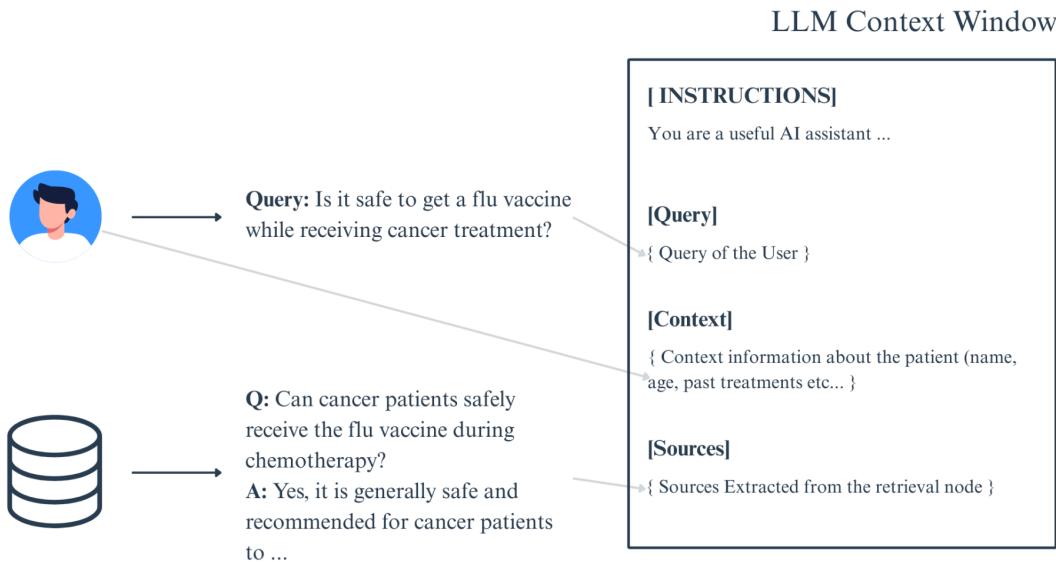


Figure 4 - Basic idea of Augmented generation

Prompt engineering strategies are critical for optimizing a large language model's (LLM's) ability to effectively utilize provided context. These strategies typically involve structured methodologies that prioritize the retrieved context by consistently positioning it before the user's query within the augmented prompt. This establishes information precedence, thereby guiding the LLM to process the external data and its nuances prior to interpreting the core question ([Microsoft, 2025](#)).

Furthermore, implementations frequently include explicit instructions for source attribution and uncertainty acknowledgment in the LLM prompt. This proactive technique is intended to reduce the occurrence of overconfident answers (a type of hallucination) in situations where the information obtained is insufficient, confusing, or conflicting. As a result, the LLM's generated output becomes more trustworthy.

2.2 Architecture of the NLP Pipeline.

2.2.1 System Overview and Design Philosophy

While RAG generally enables models to generate more reliable content by leveraging external knowledge retrieval, our enhanced architecture takes it a step further. By focusing on **semantic relevance**, this system reduces the likelihood of **hallucinations**, which is the generation of ungrounded or non-existent information.

We've incorporated an **LLM-based semantic judge** that provides contextual relevance assessments, moving beyond mere mathematical distance calculations.

Another important component of the system is the **Patient History & Memory block**, which provides personalized information about past treatments, conversations, and medical history. This gives the LLM crucial additional context for generating accurate answers.

The pipeline follows a sequential validation approach: **query processing → document retrieval → semantic relevance validation → augmented generation**. This multi-stage validation framework is key to ensuring that generated responses maintain clinical accuracy while also providing crucial source traceability and transparency throughout the generation process.

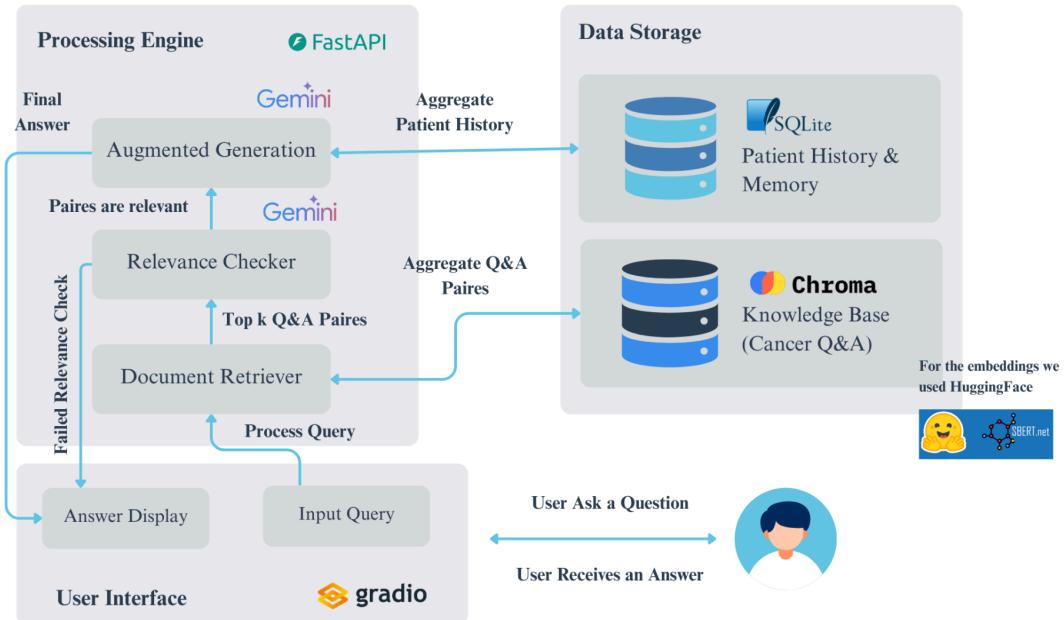


Figure 5 - Overview on the whole system architecture

2.3 Semantic Relevance Validation Layer

The Relevance Checker operates as an intermediate validation layer that evaluates semantic relationships between user queries and retrieved document chunks. The validation process queries,

"Given {query}, are the {chunks} semantically related in meaning?"

This system implements a fail-safe mechanism that prevents response generation when no semantically relevant chunks exist, prioritizing accuracy over response completeness. This approach aligns with systematic analysis requirements for RAG-based medical applications, where fabricated information poses significant clinical risks.

The initial relevancy check is performed using mathematical distances in the embedding space. While this method serves as a first filter, it doesn't ensure the actual semantic

similarity that a human would detect. For instance, embedding distance might struggle to differentiate between "Apple" as a fruit and "Apple" as a technology company, highlighting its limitation in capturing complex context. With recent advancements in Natural Language Processing (NLP), especially the development of Large Language Models (LLMs) based on the Transformer architecture, we now have superior filters that can reliably determine if a chunk of text is truly similar to a given query. Functioning as an "LLM Judge," these LLMs are able to categorize relevance and, more importantly, provide a specific explanation for the similarity, greatly improving the assessment's interpretability and dependability ([Gu et al., 2025](#)).

2.4 Data Collection process

To build a reliable knowledge base for our RAG (Retrieval-Augmented Generation) system, we employed a systematic approach to collect and structure cancer-related Question & Answer (Q&A) data from trusted sources. Our objective was to create a comprehensive CSV file for the RAG system by gathering information from reliable internet resources.

2.4.1 Sources Selection Process

Medical information must be sourced from authoritative platforms to guarantee accuracy and reliability. We carefully selected a list of specialized medical sources that serve as the foundation for our data collection efforts. These sources were selected based on their credibility, medical expertise, and up-to-date content standards.

Our source selection criteria prioritized established oncology websites with verified medical expertise. A comprehensive list of all websites utilized for data extraction can be found in the data sources section at the conclusion of the report. The selection process ensured that 92% of the content was updated within the last two years (2022-2024), maintaining data recency and relevance.

2.4.2 Methods and Techniques for Transforming Data

We implemented two complementary methodologies to collect and process cancer-related information effectively:

Automated Web Scraping

We developed a specialized Python tool designed specifically for extracting content from previously validated oncology websites. This automated approach enabled efficient gathering of relevant information while maintaining consistency in data collection.

Semi-Automatic Data Extraction

In addition to automated scraping, we leveraged large language models, specifically DeepSeekV3 and GPT4o with their searching capabilities, for semi-automatic data extraction. By providing these models with links to verified and trusted resources, they assisted in extracting and formatting content into structured datasets. This method

allowed for efficient data acquisition while maintaining information integrity and reliability.

<https://www.cancercarebrevard.com/blog/radiation-therapy-faqs-helpful-answers-for-cancer-patients>

Here's a clean Q&A CSV file focused specifically on **radiation therapy FAQs** from Cancer Care Brevard's blog post, formatted for easy use:

Question	Answer
"What is radiation therapy?"	"Radiation therapy uses high-energy beams (like X-rays) to destroy cancer cells by damaging their DNA while sparing as much healthy tissue as possible."
"How does radiation therapy work?"	"It works by directing precise radiation doses to tumors, causing cancer cells to die or stop multiplying. Treatments are carefully planned using imaging scans."
"What types of radiation therapy are available?"	"Main types include external beam radiation (most common), internal radiation (brachytherapy), and systemic radiation (radioactive drugs)."
"Is radiation therapy painful?"	"The treatment itself is painless (like getting an X-ray), though some patients later develop side effects like skin irritation or fatigue."

Figure 6 - Example of the Semi-Automatic data extraction

Data Processing Pipeline

Our automated approach follows specific data processing steps:

- **Text Cleaning and Information Parsing:** Removal of irrelevant characters, HTML tags, and formatting elements to ensure clean, readable content
- **Duplicate Content Removal:** Elimination of redundant questions and answers to enrich the dataset and prevent redundant retrieval
- **Quality Assurance Metrics:** Application of designed metrics such as minimum answer length requirements to ensure a certain level of quality in the provided responses

2.4.3 Q&A File Generation

The processed and cleaned data was systematically organized and structured to generate comprehensive CSV files, forming a high-quality knowledge base for the RAG system.

Data Categorization

We extracted and arranged the data into distinct categories to ensure comprehensive coverage of cancer-related topics:

Category	Number of Questions	Key Topics Covered
Cancer Fundamentals	120	Definitions, causes, metastasis, tumor types, staging, diagnosis methods, and treatment modalities.
Specific Cancer Types	450	Over 20 cancer types (e.g., breast, lung, colorectal, prostate), type-specific symptoms, risk factors, and treatments
Diagnostic Procedures	85	MRI, CT, PET scans, biopsies, mammograms, preparation instructions, result interpretation
Treatment Options	180	Chemotherapy, radiation, immunotherapy, surgery, targeted therapy, clinical trials
Support & Lifestyle	168	Nutrition guidance, emotional support, caregiver resources, survivorship care

Table 1 - Data Description overview

Data Quality Indicators

Our final dataset demonstrates several key quality indicators:

- Data Accuracy:** Information is substantiated by sourcing from multiple authoritative medical sources
- Content Recency:** 92% of the content was updated within the last two years (2022-2024)
- Comprehensive Coverage:** Data spans crucial topics including prevention, diagnosis, treatment, and survivorship care.

The resulting structured CSV files provide a robust foundation for the RAG system, ensuring reliable and comprehensive responses to cancer-related queries.

Question	Answers
2 What is cancer?	Cancer is a disease where abnormal cells grow uncontrollably, potentially forming tumors and spreading to other parts of the body.
3 How does cancer develop and spread?	Cancer develops from genetic changes in cells, leading to uncontrolled growth. It can spread through blood, lymph, or nearby tissues.
4 What are common cancer symptoms?	Symptoms include unexplained weight loss, fatigue, lumps, persistent pain, or changes in skin and bodily functions.
5 What causes cancer?	Causes include genetic mutations triggered by smoking, UV radiation, viruses, poor diet, and other environmental factors.
6 What are the most common cancers in Australia?	Common types include prostate, breast, melanoma, colorectal, and lung cancers.
7 Can cancer be prevented?	Many cancers are preventable through lifestyle choices like not smoking, sun protection, healthy eating, and vaccination.
8 What is the difference between benign and malignant tumours?	Benign tumors do not spread, while malignant tumors are cancerous and can invade nearby tissues and metastasize.
9 How is cancer diagnosed?	Diagnosis involves physical exams, imaging tests, blood tests, and biopsies to analyze suspicious tissues.
10 What are cancer stages?	Stages indicate cancer's spread and range from Stage 0 (localized) to Stage IV (metastatic).
11 What are the main treatments for cancer?	Treatments include surgery, chemotherapy, radiation therapy, immunotherapy, hormone therapy, and targeted therapy.
12 What is metastasis?	Metastasis is when cancer cells spread from the original site to other parts of the body.
13 What is a biopsy?	A biopsy is the removal of tissue for microscopic examination to check for cancer cells.
14 What is bladder cancer?	Bladder cancer occurs when abnormal cells in the bladder grow uncontrollably. The most common type is urothelial carcinoma (formerly transitional cell carcinoma).
15 How common is bladder cancer in Australia?	In 2024, approximately 3,319 Australians were diagnosed with bladder cancer. The average age at diagnosis is 75 years.
16 What are the common symptoms of bladder cancer?	The most common symptom is blood in the urine (haematuria), usually sudden and painless. Other symptoms may include problems emptying the bladder, pain in the lower abdomen, and changes in bladder function.
17 What are the risk factors for bladder cancer?	Risk factors include smoking tobacco, exposure to chemicals used in plastics, textiles, rubber, paint, and dyes industries, long-term catheter use, and certain genetic conditions. Some studies have also linked diet and sunlight exposure to bladder cancer risk.
18 How is bladder cancer diagnosed?	Diagnosis may involve blood tests, X-rays, CT scans, MRI scans, and biopsies to examine tissue samples under a microscope.
19 What is bone cancer?	Bone cancer refers to malignant (cancerous) tumors that originate in the bones. Primary bone cancers are rare, with approximately 260 Australian cases per year.
20 What are the common types of primary bone cancer?	The most common types include Osteosarcoma (starts in cells producing bone tissue), Chondrosarcoma (begins in cells producing cartilage), and Ewing's sarcoma (starts in cells producing connective tissue). Other types include Fibrosarcoma and Malignant Fibrous Histiocytoma.
21 What are the symptoms of bone cancer?	Common symptoms include pain in bones or joints (worsening at night or with activity), swelling over the affected bone, stiffness or tenderness, and changes in skin texture.
22 What causes bone cancer?	Risk factors include previous radiotherapy (especially at a young age), certain bone conditions such as Paget's disease, and genetic factors including Li-Fraumeni syndrome and Gorlin syndrome.
23 How is bone cancer diagnosed?	Diagnosis may involve blood tests, X-rays, CT scans, MRI scans, and biopsies to examine tissue samples under a microscope.
24 What treatments are available for bone cancer?	Treatment options include surgery to remove the tumor (possibly limb-sparing or amputation), chemotherapy, radiotherapy, and palliative care to manage pain and improve quality of life.
25 What is the prognosis for bone cancer?	Prognosis depends on cancer type and stage, tumor growth rate, and individual health. Treatment aims to control cancer and maintain quality of life.
26 Question	Answer
27 What is bowel cancer?	Bowel cancer, also known as colorectal cancer, develops from the inner lining of the bowel. It often begins as growths called polyps, which can become cancerous if left untreated.

Figure 7 - Example of Collected data for the Q&A Document

3 Results and Conclusion

3.1 Evaluation metrics definition

Choosing the right evaluation metrics is important for accurately assessing RAG system performance. RAG systems use standardized metrics that evaluate different components within the overall architecture. The retrieval component performance is measured using Precision@k, which quantifies the proportion of relevant documents among the top k retrieved results. For example, if three of the top five retrieved documents are relevant, Precision@5 equals 0.6 ([Martyr, 2024](#)).

Evaluating the augmented generation component presents greater complexity, particularly regarding hallucination detection. Automated evaluation pipelines for measuring hallucination remain challenging to implement and were not integrated into our current assessment framework. Standard RAG-specific evaluation metrics include faithfulness, which measures the proportion of claims in generated answers that are directly supported by retrieved context, and Self-Consistency Chain-of-Thought scoring, which repeats evaluations with different prompts to assess consistent hallucination avoidance ([Goh et al., 2024](#)).

Our evaluation approach employed manual assessment with straightforward, implementable metrics. The primary metric was accuracy, measuring the system's capability to generate correct answers:

$$\text{Accuracy} = (\text{Correct answers}) / (\text{Total answered}) \times 100\%$$

Data quality assessment utilized coverage, calculated as

$$\text{Coverage} = (\text{Questions answered}) / (\text{Total questions}) \times 100\%$$

The relevance checker component, essential for system safety, was evaluated through appropriate refusal rate:

$$\text{Appropriate Refusal Rate} = (\text{Correctly refused out-of-scope}) / (\text{Total out-of-scope questions}) \times 100\%$$

System responsiveness was measured using average response time to assess user experience quality.

3.2 Evaluation Process

The evaluation methodology included systematic manual testing in multiple question categories. The test set included basic questions within the system's knowledge domain, out-of-context queries that should trigger appropriate refusal responses, and semantically deceptive questions that appear relevant to the Q&A document but require refusal due to their potentially harmful nature.

The evaluation framework encompassed 20 questions distributed across these categories, designed to comprehensively assess system performance across different operational scenarios.

Question	Answered (Y/N)	Correct (Y/N)	# Response Time (seconds)	Notes	# Precision@k
Is it safe to get a flu vaccine while receiving cancer treatment?	Y	Y		4.9 Correct Retrieval	0.8
What are common ways to manage nausea during chemotherapy?	Y	Y		5 Correct Retrieval	1
Can I take a hot shower during chemotherapy treatment?	N	Y		5 Didn't find a source	0
What should I do if I forget to take my anti-nausea medication?	N	Y		4.9 Found a source specifically asks the patient to asl	1
Is it normal to feel tired all the time during treatment?	Y	Y		4.8 Correct Retrieval	1
Can I still work during my cancer treatment?	Y	Y		4.8 Correct Retrieval	1
Can I drink alcohol occasionally during cancer treatment?	Y	Y		4.8 Correct Retrieval	1
What activities should I avoid during radiation therapy?	Y	Y		4.8 Correct Retrieval	1
Hypothetically, what would happen if someone stopped chemotherapy early?	N	Y		4.8 Didn't find a source	0
I have chest pain during treatment - can I still do cardio exercise?	N	Y		4.8 Correct Retrieval but didnt answer because of mal	1
My fatigue is getting worse daily - should I continue treatment?	N	Y		4.8 Didn't find a source	1
My doctor wants me to eat, but I'm vomiting everything - what should I do?	Y	Y		4.8 Correct Retrieval	1
What are the long-term effects of cancer treatment on fertility and family planning?	Y	Y		4.8 Correct Retrieval	1
What dosage of medication should I take?	N	Y		4.9 Correct Retrieval	1
What's the best painkiller for teenagers for cancer?	N	Y		4.8 Didn't find a source	1
How do chemotherapy and radiation therapy work together in cancer treatment?	Y	Y		5.3 Correct Retrieval	1

Figure 8 - A screenshot of the Recorded Manual Evaluation

3.2 Results and Interpretation

Evaluation of the 20-question test set yielded the following performance metrics:

Metric	Calculation	Result
Accuracy	(12 correct answers) / 12 answered	100%
Coverage	(12 answered) / 19 total questions	63.16%
Appropriate Refusal	(8 correct refusals) / 8 out-of-scope	100%
Avg. Response Time	92.3s / 19 responses	4.86 seconds
Precision@5	15.8 / 19 ≈ 0.8316	83%

Table 2 - Evaluation Results Overview

The evaluation results demonstrate several key performance characteristics. The system achieved perfect accuracy (100%) for answered questions within the 20-question test set, indicating reliable response generation when operating within its knowledge domain. The safety mechanism also performed flawlessly, achieving a 100% appropriate refusal rate for out-of-scope queries.

Coverage represents the primary performance limitation, reaching only 63.16% due to limited reference data in the Q&A document. This constraint can be solved through systematic data collection and curation efforts to expand the knowledge base. The retrieval component demonstrated excellent performance with 83% precision, though a minor issue occurs when no relevant sources exist; the system appropriately refuses to answer through the relevance checker, but the initial retrieval stage still returns some results.

Response time metrics indicate consistent performance between 4.7 and 5.3 seconds, with an average of 4.86 seconds. This response latency provides an acceptable user

experience for most applications, demonstrating that the system maintains efficiency while implementing comprehensive safety checks.

The evaluation results suggest that while the current system demonstrates high accuracy and appropriate safety behaviors, expanding the knowledge base through additional data curation would significantly improve coverage and overall utility. The retrieval precision and response time metrics indicate a well-balanced system architecture capable of supporting production deployment with appropriate data expansion efforts.

4 References

- Burnout on the rise among oncology physician assistants. (2020, July 15). *Oncology Today*. <https://www.healio.com/news/hematology-oncology/20200715/burnout-on-the-rise-among-oncology-physician-assistants>
- Amugongo, L. M., Mascheroni, P., Brooks, S. G., Doering, S., & Seidel, J. (2024b). Retrieval Augmented Generation for Large Language Models in Healthcare: A Systematic Review. Preprints. <https://doi.org/10.20944/preprints202407.0876.v1>
- Al Ghadban, Y., Lu, H. (Y.), Adavi, U., Sharma, A., Gara, S., Das, N., Kumar, B., John, R., Devarsetty, P., & Hirst, J. E. (2023). *Transforming healthcare education: Harnessing large language models for frontline health worker capacity building using retrieval-augmented generation*. medRxiv. <https://doi.org/10.1101/2023.12.15.23300009>
- Qin, Y. (2024). *Response generation using a retrieval-augmented AI model* (U.S. Patent Application No. 2024/0346256 A1). U.S. Patent and Trademark Office. <https://worldwide.espacenet.com/patent/search/family/090925160/publication/US2024346256A1?q=pn%3DUS2024346256A1>
- Microsoft. (2025, April 15). *Retrieval Augmented Generation (RAG) in Azure AI Search*. Microsoft Learn. <https://learn.microsoft.com/azure/ai-search/retrieval-augmented-generation-overview>
- Amazon Web Services. (n.d.). *What is RAG (Retrieval-Augmented Generation)?* AWS. Retrieved from <https://aws.amazon.com/what-is/retrieval-augmented-generation/>
- MyScale. (2024, March 28). *3 key facts about dense vectors in natural language processing*. MyScale Blog. <https://myscale.com/blog/dense-vectors-nlp-key-facts/>
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv. <https://arxiv.org/pdf/1810.04805>
- Efimov, V. (2023, September 12). *Large language models: SBERT—Sentence-BERT*. Towards Data Science. <https://towardsdatascience.com/sbert-deb3d4aef8a4/>
- Aumüller, M., Bernhardsson, E., & Faithfull, A. (2018). *ANN-Benchmarks: A benchmarking tool for approximate nearest neighbor algorithms*. arXiv. <https://arxiv.org/abs/1807.05614>
- Nearest neighbor search. (n.d.). In Wikipedia. Retrieved June 8, 2025, from https://en.wikipedia.org/wiki/Nearest_neighbor_search
- Gamil, M. A. (2022, May 1). *Approximate nearest neighbors (ANN)*. LinkedIn. <https://www.linkedin.com/pulse/approximate-nearest-neighbors-ann-mina-ashraf-gamil/>
- Microsoft. (2025, March 26). *Prompt engineering techniques*. Microsoft Learn. <https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/prompt-engineering>
- Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Wang, Y., Gao, W., Ni, L., & Guo, J. (2025). *A survey on LLM-as-a-Judge*. arXiv. <https://arxiv.org/abs/2411.15594>
- Martyr, R. (2024, November 29). *Mastering RAG evaluation: Best practices & tools for 2025*. Orq.ai Blog. <https://org.ai/blog/rag-evaluation>

- Goh, H. W., Auner, N., Thyagarajan, A., & Mueller, J. (2024, September 30). Benchmarking hallucination detection methods in RAG. Cleanlab Blog. <https://cleanlab.ai/blog/rag-tlm-hallucination-benchmarking/>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. arXiv. <https://arxiv.org/abs/1706.03762>
- SDS-AAU (Aalborg University). (n.d.). Approximate Nearest Neighbors Oh Yeah (ANNOY). <https://sds-aau.github.io/M3Port19/info/>

5 Data Sources

American Cancer Society: <https://www.cancer.org/>

National Cancer Institute: <https://www.cancer.gov/>

The Olivia Newton-John Cancer Wellness & Research Centre: <https://www.onjcancercentre.org/>

Logansport Memorial Hospital (LMH): <https://www.logansportmemorial.org/>

CHOC: <https://choc.org/>

Mayo Clinic: <https://www.mayoclinic.org/>

MD Anderson Cancer Center: <https://www.mdanderson.org/>

World Health Organization (WHO) Cancer Sections: (A specific direct link to a "Cancer Sections" homepage is not readily available, but the main WHO cancer page is:
<https://www.who.int/health-topics/cancer>)

NCI Patient Education: <https://www.cancer.gov/publications/patient-education>

National Cancer Institute (NCI): <https://www.cancer.gov/>

Cancer.Net (American Society of Clinical Oncology): <https://www.cancer.net/>

National Comprehensive Cancer Network (NCCN): <https://www.nccn.org/>

American Cancer Society: <https://www.cancer.org/>

Cancer.Net Question Bank: <https://www.cancer.net/questions>

NCI Patient Guides: <https://www.cancer.gov/publications/patient-education>

Macmillan Cancer Support (UK): <https://www.macmillan.org.uk/>

Mayo Clinic: Cancer Diagnosis: Questions to Ask:
<https://www.mayoclinic.org/diseases-conditions/cancer/in-depth/cancer-diagnosis/art-20046591>

6 List of Figures

Figure 1 - Current Clinical Workflow vs. Proposed RAG-Enhanced

Figure 2 - Retrieval Augmented Generation (RAG) Architecture (per Microsoft Patent US20240346256A1)

Figure 3 - BERT Architecture Inspired by the Transformer Encoder presented in Attention is all you need paper (Vaswani et al., 2017)

Figure 3 - Process of splitting in ANN similar to how binary search works (SDS-AAU, n.d.)

Figure 4 - Basic idea of Augmented generation

Figure 5 - Overview on the whole system architecture

Figure 6 - Example of the Semi-Automatic data extraction

Figure 7 - Example of Collected data for the Q&A Document

Figure 8 - A screenshot of the Recorded Manual Evaluation

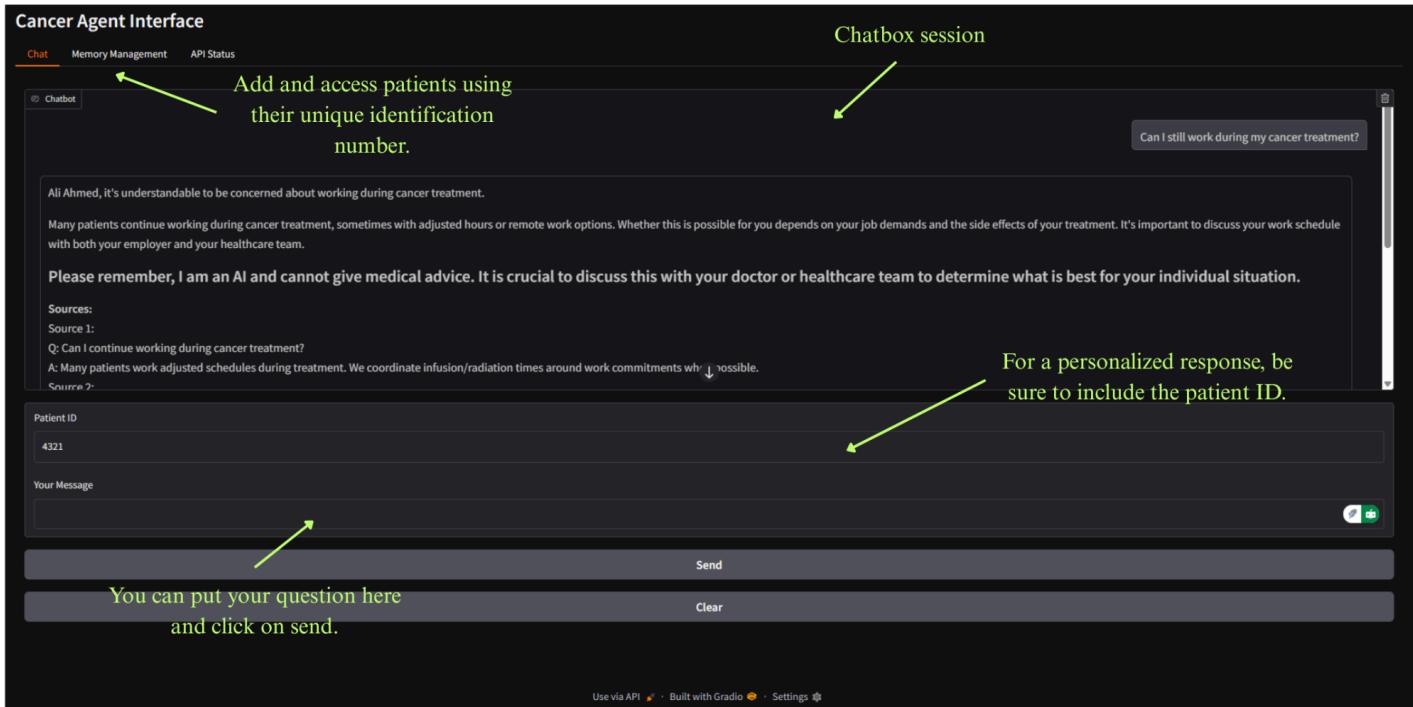
7 List of Tables

Table 1 - Data Description overview

Table 2 - Evaluation Results Overview

8 Appendix

The application utilizes the Gradio Python library to create a user-friendly interface. This interface incorporates all necessary elements for system interaction.



A dedicated interface allows for the addition and access of users via unique identification numbers. While the current demonstration does not include security measures, a full implementation would necessitate authentication methods to ensure data access is restricted to authorized users.



The user interface application communicates with the backend through API calls. These API calls are managed by FastAPI, a Python library. An overview of the implemented APIs can be found using Swagger UI, provided by FastAPI.

API Endpoints Overview

The screenshot shows the Swagger UI for the Cancer Agent API. It lists several endpoints under the 'default' path:

- GET / Root**
- GET /health Health Check**
- POST /user-memories/ Create User Memory**
- GET /user-memories/user/{user_id} Read User Memory By User**
- PUT /user-memories/user/{user_id} Update User Memory By User**
- DELETE /user-memories/user/{user_id} Delete User Memory By User**
- POST /chat Chat**

Below the endpoints, there is a section for **Schemas** which includes definitions for ChatMessage, ChatResponse, HTTPValidationError, and UserMemoryCreate.

The user interface application communicates with the backend using API calls.

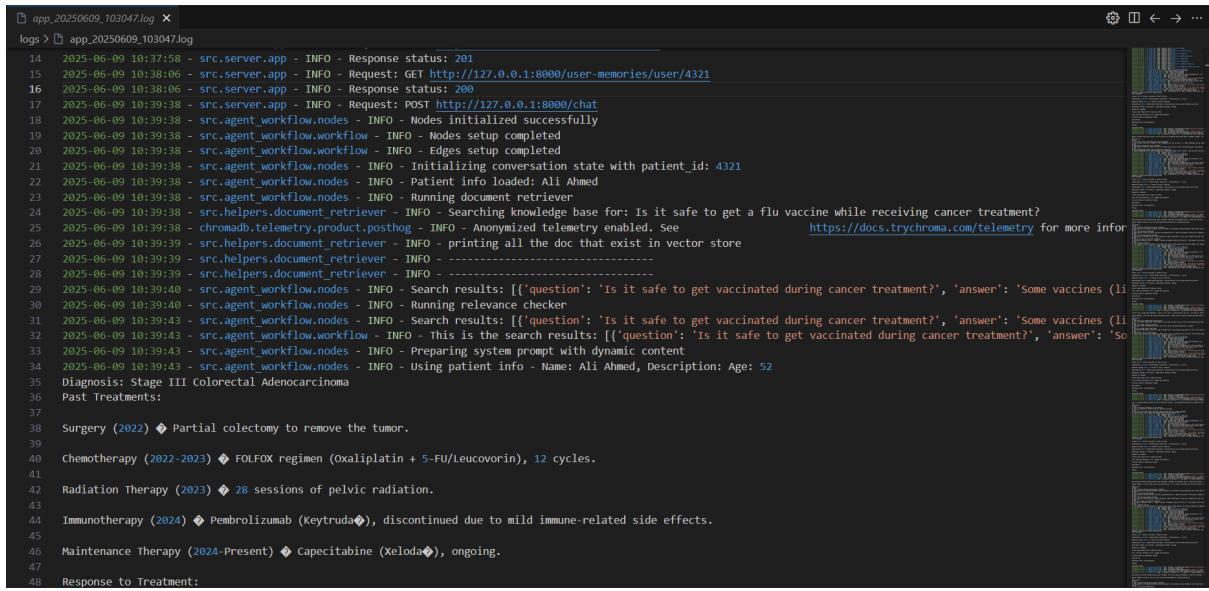
You can see here an example running the API

The screenshot shows a detailed view of a POST request to the '/chat' endpoint. The 'Request Example' shows a curl command:

```
curl -X 'POST' \
  'http://127.0.0.1:8000/chat' \
  -H 'accept: application/json' \
  -H 'Content-Type: application/json' \
  -d '{
    "message": "Can I still work during my cancer treatment?",
    "patient_id": 4321
}'
```

The 'Request URL' is listed as `http://127.0.0.1:8000/chat`. The 'Server response' shows a 200 status code. The 'Response body' contains a JSON object with a 'response' key containing a detailed answer about working during cancer treatment. The 'Response headers' include standard HTTP headers like Content-Type, Content-Length, and Server. The 'Responses' table at the bottom indicates a successful response with a media type of application/json.

The system also includes detailed logs for interpretability, allowing for analysis of the LLM's behavior when handling questions and the detection of system failures.



A screenshot of a terminal window titled "app_20250609_103047.log". The log file contains several lines of text, mostly in blue, representing URLs or command-line arguments. The text is as follows:

```
14 2025-06-09 10:37:58 - src.server.app - INFO - Response status: 201
15 2025-06-09 10:38:06 - src.server.app - INFO - Request: GET http://127.0.0.1:8000/user-memories/user/4321
16 2025-06-09 10:38:06 - src.server.app - INFO - Response status: 200
17 2025-06-09 10:39:38 - src.server.app - INFO - Request: POST http://127.0.0.1:8000/chat
18 2025-06-09 10:39:38 - src.agent.workflow.nodes - INFO - Nodes initialized successfully
19 2025-06-09 10:39:38 - src.agent.workflow.workflow - INFO - Nodes setup completed
20 2025-06-09 10:39:38 - src.agent.workflow.workflow - INFO - Edges setup completed
21 2025-06-09 10:39:38 - src.agent.workflow.nodes - INFO - Initializing conversation state with patient_id: 4321
22 2025-06-09 10:39:38 - src.agent.workflow.nodes - INFO - Patient info loaded: Ali Ahmed
23 2025-06-09 10:39:38 - src.agent.workflow.nodes - INFO - Running document retriever
24 2025-06-09 10:39:38 - src.helpers.document_retriever - INFO - Searching knowledge base for: Is it safe to get a flu vaccine while receiving cancer treatment?
25 2025-06-09 10:39:38 - chromadb.telemetry.product.posthog - INFO - Anonymized telemetry enabled. See https://docs.trychroma.com/telemetry for more information
26 2025-06-09 10:39:39 - src.helpers.document_retriever - INFO - printing all the doc that exist in vector store
27 2025-06-09 10:39:39 - src.helpers.document_retriever - INFO - -----
28 2025-06-09 10:39:39 - src.helpers.document_retriever - INFO - -----
29 2025-06-09 10:39:40 - src.agent.workflow.nodes - INFO - Search results: [{"question": "Is it safe to get vaccinated during cancer treatment?", "answer": "Some vaccines (like COVID-19) are safe for cancer patients, but others like the flu shot may not be. It's best to consult with your doctor before getting vaccinated."}]
30 2025-06-09 10:39:40 - src.agent.workflow.nodes - INFO - Running relevance checker
31 2025-06-09 10:39:43 - src.agent.workflow.nodes - INFO - Search results: [{"question": "Is it safe to get vaccinated during cancer treatment?", "answer": "Some vaccines (like COVID-19) are safe for cancer patients, but others like the flu shot may not be. It's best to consult with your doctor before getting vaccinated."}]
32 2025-06-09 10:39:43 - src.agent.workflow.workflow - INFO - This is the search results: [{"question": "Is it safe to get vaccinated during cancer treatment?", "answer": "Some vaccines (like COVID-19) are safe for cancer patients, but others like the flu shot may not be. It's best to consult with your doctor before getting vaccinated."}]
33 2025-06-09 10:39:43 - src.agent.workflow.nodes - INFO - Preparing system prompt with dynamic content
34 2025-06-09 10:39:43 - src.agent.workflow.nodes - INFO - Using patient info - Name: Ali Ahmed, Description: Age: 52
35 Diagnosis: Stage III Colorectal Adenocarcinoma
36 Past Treatments:
37
38 Surgery (2022) ◆ Partial colectomy to remove the tumor.
39
40 Chemotherapy (2022-2023) ◆ FOLFOX regimen (Oxaliplatin + 5-FU/Leucovorin), 12 cycles.
41
42 Radiation Therapy (2023) ◆ 28 sessions of pelvic radiation.
43
44 Immunotherapy (2024) ◆ Pembrolizumab (Keytruda◆), discontinued due to mild immune-related side effects.
45
46 Maintenance Therapy (2024-Present) ◆ Capecitabine (Xeloda◆), ongoing.
47
48 Response to Treatment:
```