PROBLEM STATEMENT

PS-04 : Introduction to GenAI and Simple LLM Inference on CPU and fine-tuning of LLM Model to create a Custom Chatbot

This project involves exploring Generative AI by performing LLM inference on a CPU and fine-tuning a pre-trained LLM model to develop a custom chatbot. Key tasks include managing large model files, executing inference processes, and applying fine-tuning techniques using Intel AI Tools. The goal is to gain practical experience in these areas, focusing on technical challenges and solutions.

OUR SOLUTION

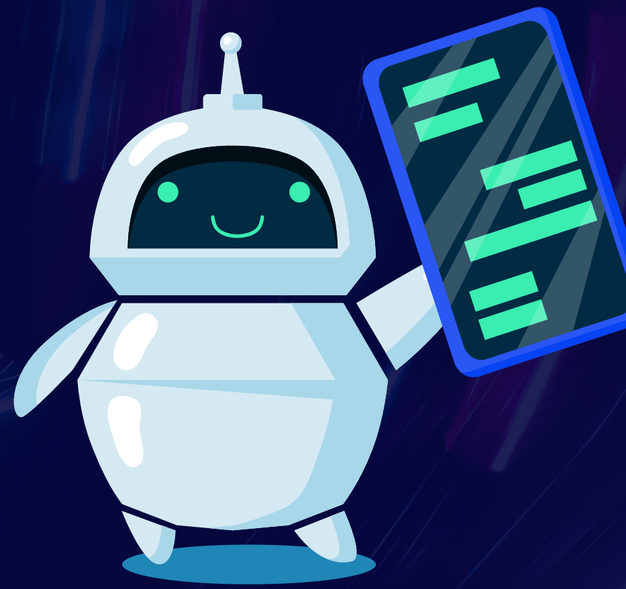
- **AI-Powered Chatbot:** Developed an advanced AI-powered chatbot tailored to provide accurate and detailed responses to cancer-related queries, ensuring users receive precise information promptly.
- **Personalized Assistance:** Offers personalized responses that cater to individual user queries, delivering tailored assistance to meet specific needs and preferences effectively.
- **Lightweight LLM:** Implemented a streamlined large language model (LLM) optimized for efficiency and effectiveness in handling oncology-related queries, ensuring smooth operation without compromising performance.
- **Only CPU Required:** Engineered the chatbot to operate seamlessly on CPUs, ensuring cost-effectiveness and ease of deployment without the need for specialized hardware, thereby enhancing accessibility and usability.
- **Data in PDF Format:** Utilizes advanced techniques to extract and utilize information from PDF-format oncology resources, simplifying the process of accessing and integrating authoritative data sources into the chatbot.
- **Utilizes Authoritative Sources:** Integrates insights from multiple reputable oncology references, ensuring that responses are comprehensive, accurate, and up-to-date with the latest medical knowledge and practices.

FEATURES

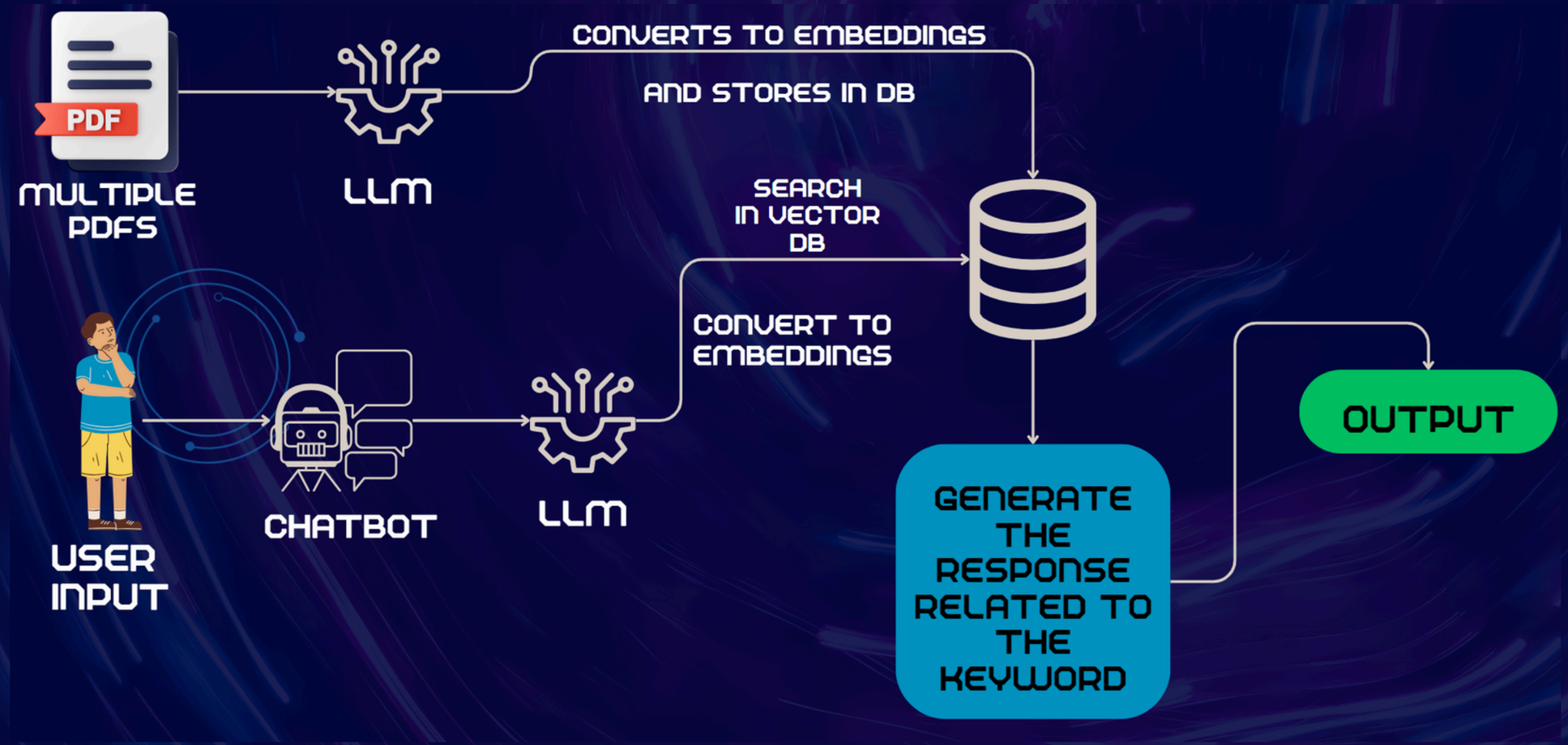
- **Advanced Natural Language Understanding:** Incorporates advanced natural language understanding (NLU) techniques to accurately interpret and comprehend user questions about cancer.
- **Comprehensive Knowledge Base:** Leveraging multiple authoritative oncology books, the chatbot's knowledge base is extensive and ensures comprehensive coverage of cancer-related information.
- **Interface:** Designed with usability in mind, the chatbot features an intuitive interface that enhances user interaction.
- **Personalized Responses:** To meet individual user needs, the chatbot generates personalized responses. By understanding the context of queries, the chatbot delivers tailored information that addresses specific concerns effectively.
- **Continuous Assistance:** The chatbot provides continuous access to information about cancer, ensuring users can obtain timely assistance whenever required.

PROCESS FLOW

- **User Query Submission:** Users submit cancer-related questions through the chatbot's web interface. Users can type in their queries about cancer, whether they are seeking information about symptoms, treatments, or general cancer knowledge.
- **NLP Processing:** The chatbot uses NLP techniques to analyze and interpret the submitted queries. The NLP engine breaks down the user's query into understandable parts, identifies the key components, and prepares the data for further processing.
- **Information Retrieval:** After the query is processed by the NLP engine, the system retrieves relevant information from its knowledge base. The knowledge base is built using Chroma as the vector database, which stores and manages the embeddings of the information extracted from the dataset.
- **Response Generation:** With the relevant information retrieved, the chatbot formulates a detailed and accurate response. This step utilizes BGE Embeddings and the fine-tuned Neural Chat v3 model by Intel. BGE Embeddings ensure the response is coherent and contextually appropriate, while the fine-tuned Neural Chat v3 model generates informative responses aligned with the user’s query.
- **Follow-Up Interaction:** The chatbot supports follow-up interactions to provide additional details or clarification, ensuring comprehensive user support.



ARCHITECTIRE DIAGRAM



TECH STACK

Frontend Development: Developed using HTML, CSS, and JavaScript for a responsive and user-friendly interface.

Backend Development: Flask framework handles server-side logic and API management.

Vector Database: Chroma is used to manage embeddings and enable efficient data retrieval.

Embedding: BGE Embeddings are utilized for encoding data, facilitating accurate information retrieval.

LLM: Neural Chat v3 by Intel, fine-tuned for domain-specific responses.

Frameworks: Langchain and CTransformers are employed to integrate and manage the chatbot's various AI components.

USER INTERFACE

Our chatbot's interface includes a user-friendly input box for entering cancer-related queries, complemented by a straightforward submit button for quick access to accurate information from authoritative sources. Below is a screenshot showcasing the interface.

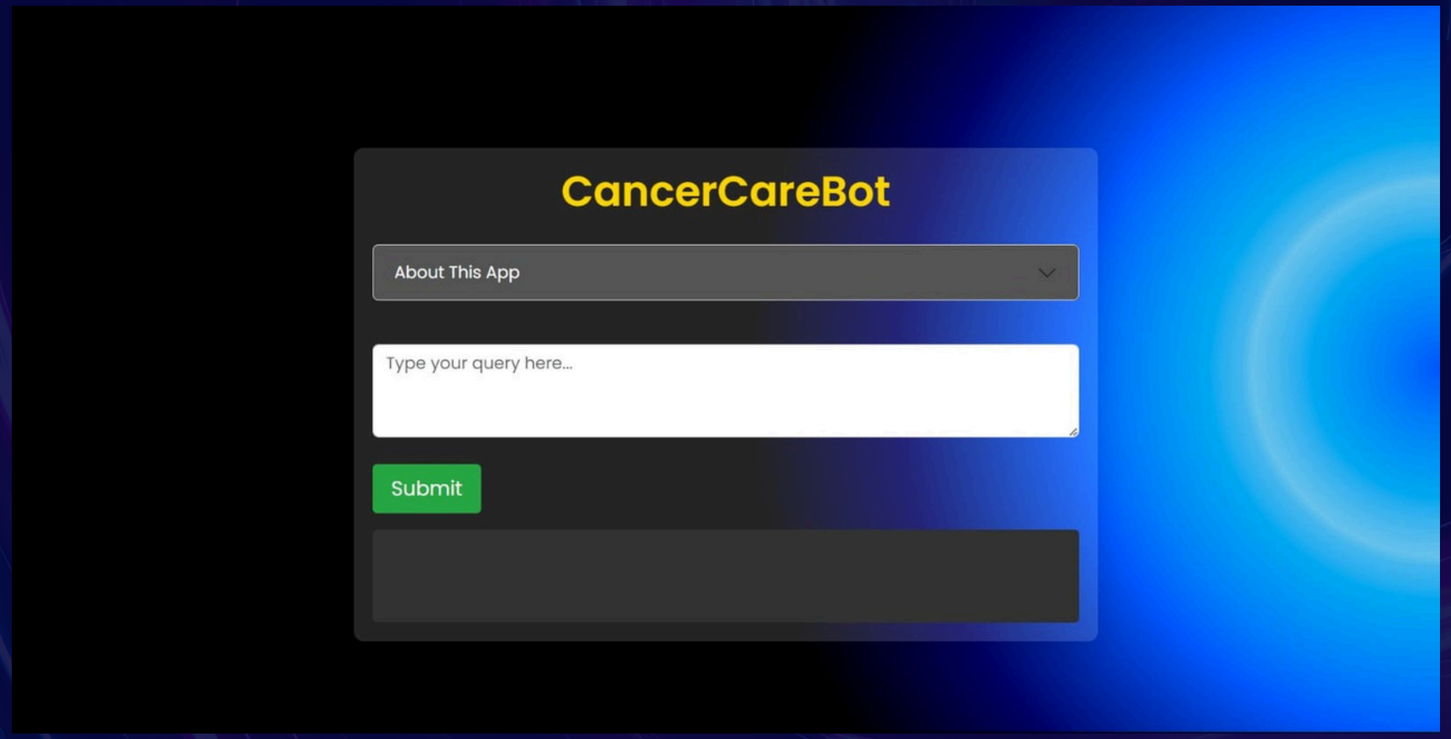


Figure 1: User Interface Screenshot

SAMPLE I/O

Below is one set of screenshots demonstrating sample interaction with our chatbot:

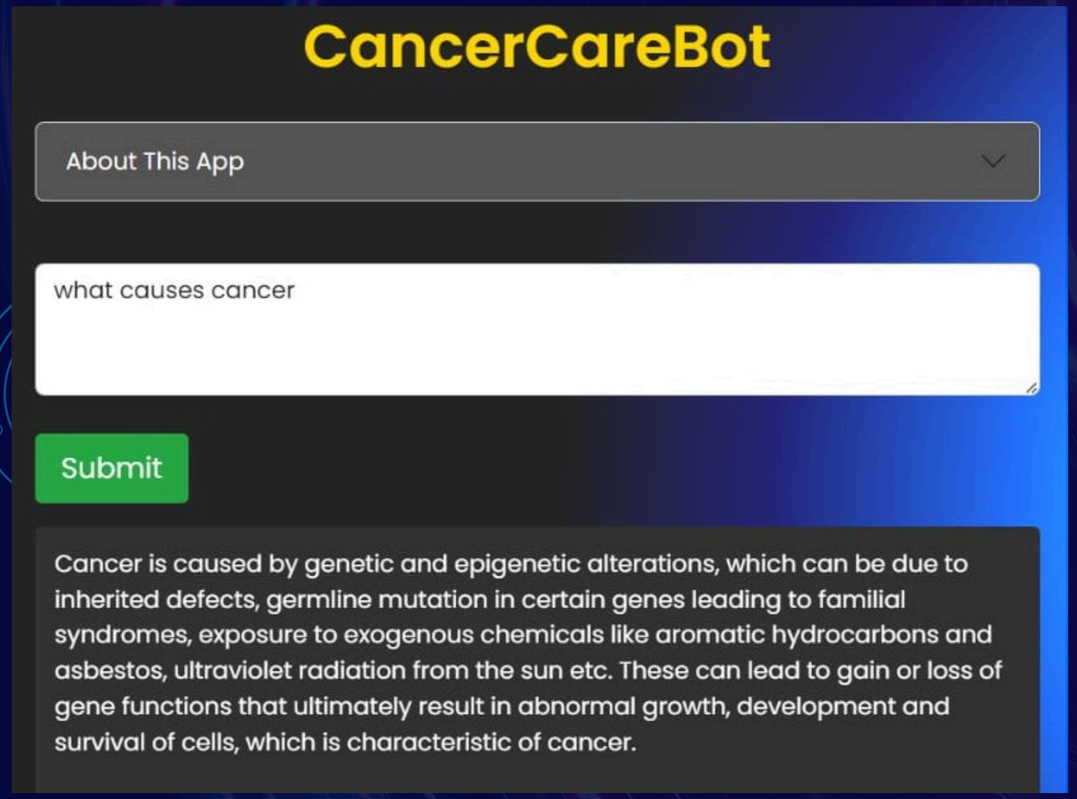


Figure 2: User Query and Chatbot Response

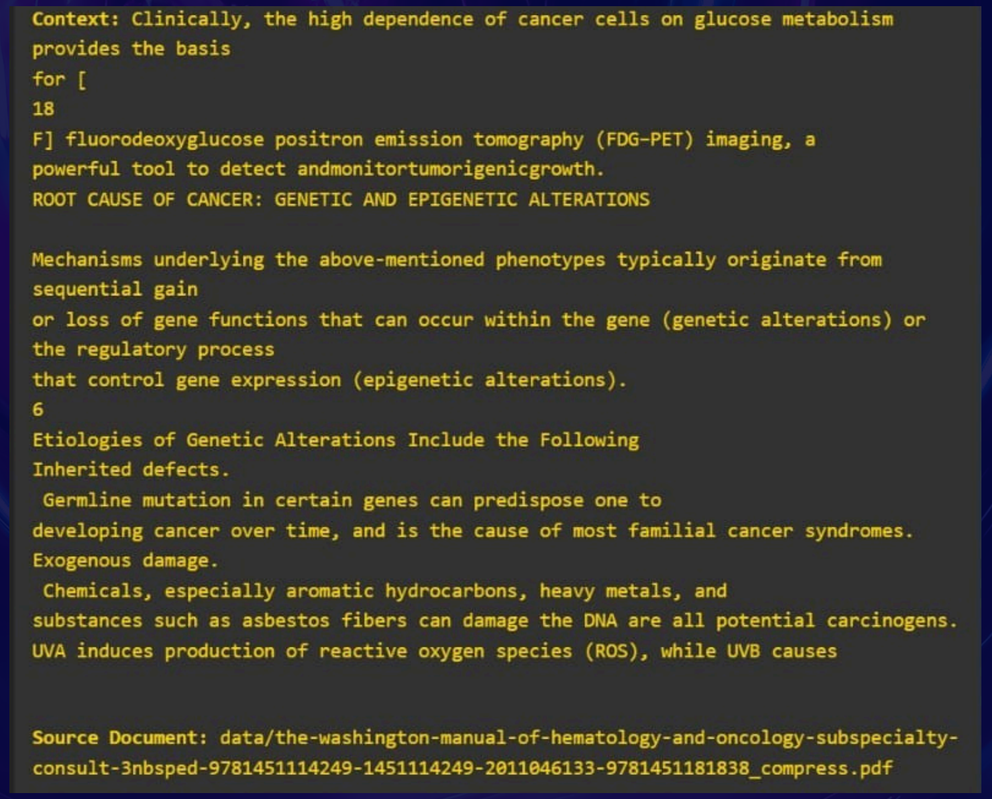


Figure 3: Context and Source Information for Query 1

SAMPLE I/O CONTINUED

Below is another set of screenshots demonstrating sample interaction with our chatbot:

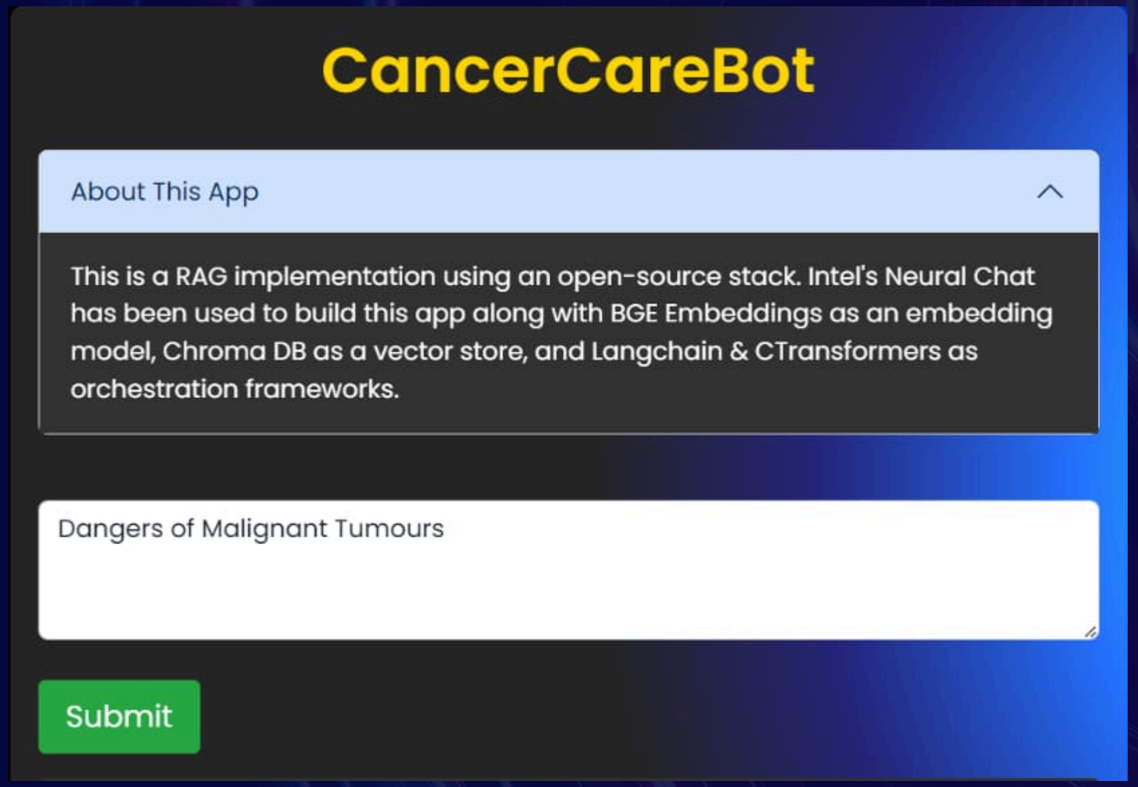


Figure 4: User Query

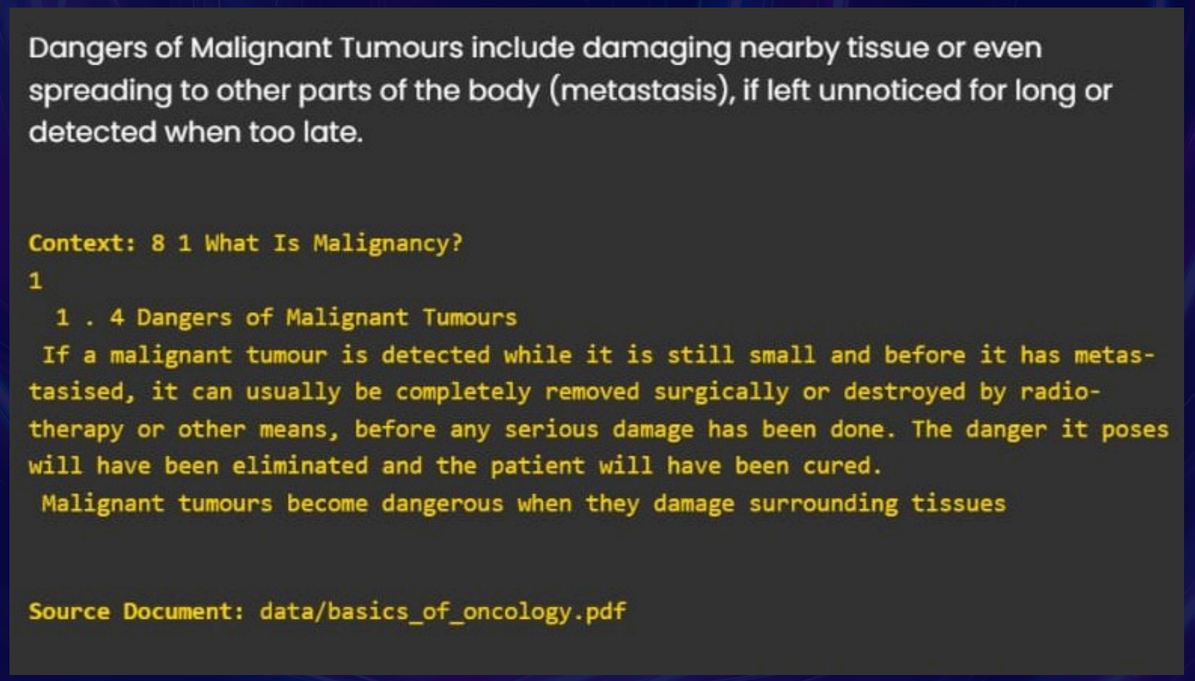


Figure 5: Chatbot response, Context and Source Information for Query 2

TRAINING TIME

Training our AI-powered chatbot is a critical aspect of its development. Here's a detailed overview of the training process and the estimated time required for our chatbot:

- **Training Process:** Our chatbot undergoes extensive training using advanced machine learning techniques to understand and respond to cancer-related queries.
- **Time Required:** The training process is estimated to take approximately 2-3 hours. This duration is influenced by factors such as the size of the dataset, the complexity of the model architecture (like the Neural Chat v3 by Intel), and the computing resources available.
- **Response Time:** Once trained, the chatbot typically responds to user queries within [30 seconds to 60 seconds], depending on the complexity and length of the query. This ensures quick and efficient interactions with users.

INSIGHTS AND LEARNINGS

- **Integration of AI in Healthcare:** We have learned the significant potential of AI in transforming healthcare by providing accurate and timely information to patients and caregivers.
- **Importance of Reliable Data Sources:** The quality of the chatbot's responses heavily relies on the accuracy and comprehensiveness of the data sources, emphasizing the need for utilizing authoritative and well-documented medical references.
- **Challenges in NLP Implementation:** Implementing NLP techniques and fine-tuning models to understand and generate relevant responses required careful handling of domain-specific language and concepts.
- **Technical Skills Enhancement:** The project enhanced our technical skills in using advanced AI tools and frameworks, such as Chroma for data retrieval and Flask for backend development.
- **User-Centric Design:** Designing a user-friendly interface that ensures accessibility and ease of use is crucial for the successful adoption of AI-driven solutions in healthcare.
- **Collaborative Effort:** Effective teamwork and division of responsibilities were key to efficiently managing the project's various components and achieving our objectives.

TEAM MEMBERS & CONTRIBUTIONS

- **Aditya Kulkarni : Project Lead**
  Oversaw the entire project, coordinated team efforts, handled data extraction and preprocessing, and integrated all components of the chatbot.
- **Kashish Aswani : NLP and Model Integration Specialist**
  Implemented NLP techniques and fine-tuned the Neural Chat v3 model for accurate query responses.
- **Rasika Rakhewar : Backend Developer**
  Developed the backend using Flask and integrated the Chroma vector database for data retrieval.
- **Nagesh Ballurkar : Frontend Developer**
  Designed and developed the user-friendly interface using HTML, CSS, and JavaScript.
- **Pallavi Laglludkar : Deployment and Testing Specialist**
  Conducted testing, managed deployment readiness, and handled troubleshooting and debugging.

CONCLUSION

- Demonstrates the application of AI for providing reliable, cancer-related information sourced from multiple authoritative oncology books.
- Implements a user-friendly interface and efficient data retrieval for quick and easy access to information.
- Enhances patient care by offering immediate, accurate information, reducing anxiety, and supporting informed decision-making.
- Provides high-quality, contextually relevant responses tailored to user needs, increasing accessibility to reliable cancer information.
- Promotes self-education and empowerment among patients and facilitates better communication with healthcare providers.