SURGICAL INFORMATION ASSISTANT: A TECHNICAL REPORT ON AN AGENTIC INFORMATION RETRIEVAL SYSTEM FOR SURGICAL INFORMATION

Kiran Bhattacharyya
SKA Labs
Atlanta, GA, USA
kiran.bhattacharyya@ska-labs.com

ABSTRACT

This technical report introduces the Surgical Information Assistant, a tool designed to improve access and utilization of surgical information. Using extensive documentation from the Vanderbilt University Open Manual of Surgery for Resource-limited Settings and employing an advanced agentic retrieval system, this assistant provides prompt, precise, and context-sensitive responses to surgical inquiries. The system is also able to access a supplementary source of general information if the user's answer is not available in the reference documents ensuring a robust and comprehensive information retrieval process. This report details the motivation, implementation, and evaluation of the Surgical Information Assistant, highlighting its potential to enhance access to vital surgical knowledge and improve medical education [1, 2]. The results demonstrate that the agentic retrieval system designed here with a small language model (llama3.2-3b) outperforms the general knowledge base of large models (GPT-4o) and other existing retrieval systems built with larger models (llama3.1-8b).

***K*eywords** surgical information · agentic retrieval system · information retrieval · medical education

1 Introduction

The field of surgery is characterized by vast amounts of critical information that medical professionals must access quickly and accurately. The Surgical Information Assistant aims to bridge the gap between this extensive medical documentation and its immediate accessibility [3, 4]. By providing accurate, context-sensitive and on-demand responses to surgical queries, the system serves as an invaluable resource for medical professionals, students, and researchers [5].

1.1 Key benefits

1. Rapid access to surgical information
2. Improved learning experience for students and medical professionals
3. Synthesis of information across multiple verified documents
4. Fallback mechanism using supplementary source (Wikipedia used here), if needed, for comprehensive coverage

1.2 Project scope

While the project primarily uses the Open Manual of Surgery in Resource-limited Settings (OMSRS) as its source documentation, this project is not officially associated with or endorsed by the OMSRS. The system is designed to build on OMSRS within the scope of the Creative Commons licensing structure.

All code and data used for this manuscript is available at this GitHub repository.

2 System architecture

The system architecture of the Surgical Information Assistant is designed to offer precise and timely information to users. This section outlines the key components and processes underlying the assistant's functionality, emphasizing its agentic retrieval system, asynchronous and sequential processing capabilities, and Wikipedia fallback mechanism.

2.1 DeRetSyn: an agentic retrieval system

The core of the Surgical Information Assistant is its sophisticated agentic retrieval system, which employs a multi-step process to provide accurate and relevant information [6, 7, 8]:

1. **Decomposition**: The user's query is analyzed and processed to extract sub-questions.
2. **Retrieval**: Each sub-question is used to search a pre-built FAISS (Facebook AI Similarity Search) index containing vector representations of the surgical manual content.
3. **Synthesis**: An LLM attempts to answer the original question based on the sub-questions and their answers.
4. **Iteration**: If a satisfactory answer is not generated, the system creates new sub-questions and repeats the process [9].
5. **Best Effort**: After three iterations, if no satisfactory answer is found, the system is referred to Wikipedia as a secondary source.

The agentic retrieval system developed here is named DeRetSyn (after Decompose, Retrieve, and Synthesize).

2.2 Asynchronous and sequential processing

The DeRetSyn is designed to run parts of retrieval process asynchronously, while other parts are executed sequentially. This hybrid approach allows for efficient processing of multiple sub-questions and retrieval tasks synchronously while maintaining the logical flow of the question-answering process with checks and iterations.

2.3 Wikipedia fallback mechanism

When the primary surgical database doesn't yield satisfactory results, the assistant employs a dual-search approach using Wikipedia [10]:

1. **Fast Search:** A rapid search is performed using ColBERTv2 on a pre-indexed Wikipedia abstract dataset.
2. **Slow Search:** A more comprehensive search is conducted using the Wikipedia Python library to retrieve full page content.
3. **Integration:** Results from both searches are combined with information from primary surgical texts to provide a comprehensive answer.

The architecture of the Surgical Information Assistant, with its robust retrieval and processing strategies, ensures that users receive accurate answers efficiently. By leveraging both primary and secondary information sources, the system maintains a high standard of reliability.

3 Implementation details

3.1 Core Components

The Surgical Information Assistant is implemented in Python and consists of several key components:

1. **FAISS Index**: A pre-built index for efficient similarity-based retrieval of surgical information.
2. **Language Models**: Utilizes advanced language models for query decomposition, answer synthesis, and evaluation.
3. **Streamlit Interface**: Provides a user-friendly web interface for interacting with the assistant.
4. **Asynchronous Processing**: Implements asynchronous functions for efficient handling of multiple retrieval and processing tasks.

3.2 QA Dataset Generation

To validate DeRetSyn, a QA dataset was created using the following steps (as detailed in 'extract_qa.py'):

1. **Text Extraction**: PDFs from the surgical manual are processed to extract text content.
2. **Chunking**: The extracted text is divided into manageable chunks.
3. **QA Pair Generation**: An LLM is used to generate question-answer pairs from each chunk [11].
4. **Asynchronous Processing**: The system can generate QA pairs asynchronously to improve efficiency.

4 Evaluation

4.1 Dataset Creation and Curation

To rigorously evaluate our surgical information retrieval system, we first created a comprehensive question-answer dataset derived from the Open Manual of Surgery in Resource-Limited Settings. The dataset creation process involved:

1. **Question Generation**: We programmatically generated using a LLM over 16,000 question-answer pairs during the document embedding process, covering a wide range of surgical topics, procedures, and techniques contained within the source documents. The context used to generate the question-answer pairs were retained for quality review.
2. **Quality Review**: The generated question-answer pairs underwent both automated filtering and manual review to ensure relevance, clarity, and faithfulness to the original context from which they were generated. Questions that were ambiguous or poorly formulated were eliminated.
3. **Dataset Refinement**: After the quality review process, the dataset was reduced to 14,529 high-quality question-answer pairs, which served as our evaluation benchmark. Note that all question-answer pairs were not reviewed manually due to resource limitations.

4.2 Evaluation Methodology

We conducted a series of comparative evaluations to characterize the performance of different retrieval-augmented generation (RAG) approaches and the general knowledge of language models. Our evaluation framework was designed to answer two key research questions: 1) Can smaller language models with access to reference data match or exceed the performance of larger models without such access? 2) How do different RAG paradigms compare in terms of accuracy when answering surgical questions? [2, 12, 13]

For each evaluation, we measured the accuracy of each of the RAG system's responses against the ground truth answers in our curated dataset. The generated response was compared to the known answer with an LLM (Mistral Small 3.1) to determine whether the generated answer was correct.

4.3 Evaluation Results

Our evaluation compared six different configurations, including models with and without RAG capabilities, and different RAG paradigms:

Model Configuration	**Accuracy (%)**
Llama 3.1-8B without RAG	28.7
GPT-4o without RAG	42.5
Llama 3.1-8B with Vanilla RAG	52.5
Llama 3.2-3B with Vanilla RAG	53.8
Llama 3.2-3B with ReACT RAG	53.5
Llama 3.2-3B with DeRetSyn RAG	**63.0**

Table 1: Accuracy comparison of different model configurations on the surgical information retrieval task

4.4 Analysis and Discussion

The evaluation results reveal several important insights:

1. **RAG Effectiveness**: All RAG-enhanced models significantly outperformed their non-RAG counterparts. Llama 3.1-8B showed a dramatic improvement from 28.7% to 52.5% accuracy when augmented with vanilla RAG, demonstrating the critical importance of retrieval for domain-specific tasks.
2. **Model Size vs. RAG**: Notably, the smaller Llama 3.2-3B model with RAG (53.8%) outperformed the larger GPT-4o model without RAG (42.5%). This confirms our hypothesis that smaller models with access to relevant reference data can exceed the performance of larger models relying solely on parametric knowledge [1, 14].
3. **RAG Paradigm Comparison**: Among the RAG approaches, our DeRetSyn (Decompose-Retrieve-Synthesize) system achieved the highest accuracy at 63%, substantially outperforming both vanilla RAG (53.8%) and ReACT RAG (53.5%) using the same base model.
4. **Efficiency Tradeoffs**: While the DeRetSyn system demonstrated superior accuracy, it requires multiple LLM calls—some of which cannot be executed asynchronously—resulting in longer response generation times and increased computational requirements compared to simpler RAG approaches implemented with the same language model [15, 7].

4.5 Computational Considerations

The DeRetSyn system, while more accurate, incurs additional computational costs due to its multi-step approach:

1. **Question Decomposition**: The system first breaks down complex queries into simpler sub-questions.
2. **Parallel Retrieval**: Where possible, retrieval operations for sub-questions are executed asynchronously.
3. **Information Synthesis**: The final step combines retrieved information to generate a comprehensive answer.

This process requires multiple LLM invocations, increasing both latency and computational requirements. However, the significant accuracy improvements (9.2 percentage points over vanilla RAG) may justify these costs in applications where answer quality is paramount, such as medical information retrieval.

5 Conclusion

The combined capabilities of smaller, more efficient language models and the Surgical Information Assistant highlight the potential for serving on-device answers to medical questions that may contain sensitive health information. Our evaluation shows that models like Llama 3.2-3B can deliver competitive performance on specialized tasks through advanced retrieval algorithms, such as the DeRetSyn RAG approach, making them suitable for resource-constrained environments [5, 9]. Additionally, the Surgical Information Assistant exemplifies the power of agentic retrieval systems in accessing and synthesizing critical surgical information. By decomposing complex queries and performing tasks asynchronously, it provides efficient and accurate information retrieval in the domain of surgical knowledge. These advancements underline the importance of AI-powered tools in ensuring immediate access to relevant medical information, paving the way for future intelligent medical systems that can localize data and models to reduce privacy concerns [15].

References

[1] K Singhal and et al. Retrieval-augmented language models for clinical medicine. *Nature*, 620:282–289, 2023.

[2] Yujia Wu and et al. Mirage: Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*, 2024.

[3] J Park and et al. Enhancing medical ai with retrieval-augmented generation. *Frontiers in AI*, 2024. PMC12059965.

[4] YJ Suh and et al. Evaluation of a context-aware chatbot using retrieval-augmented generation for osteonecrosis qa. *Journal of Oral and Maxillofacial Surgery*, 2024.

[5] T Nguyen and et al. Systematic analysis of rag-based medical qa systems in low-resource settings. *Information*, 6(4):116, 2023.

[6] Lu Chen and et al. Agentic retrieval-augmented generation: A survey. *arXiv preprint arXiv:2501.09136*, 2024.

[7] Jong Lee and et al. Surgraw: Structured multi-agent workflow with chain of thought reasoning for robotic-assisted surgery qa. *arXiv preprint arXiv:2503.10265*, 2024.

[8] Hao Zhang and et al. Discuss-rag: Enhancing retrieval-augmented generation via agent-led multi-turn reasoning for medical qa. *arXiv preprint arXiv:2504.21252*, 2024.

[9] Ryan Ahsan. Clinical reasoning with rag: Building a local medical qa system, 2023. https://medium.com/@ahsan.new252/clinical-reasoning-with-rag.

[10] Matillion. Agentic rag: Enhancing ai workflows, 2024. https://www.matillion.com/learn/blog/agentic-rag.

[11] Saurav Joshi. Retrieval-augmented generation for medical qa with llama-2–7b, 2023. https://medium.com/@sauravjoshi23/retrieval-augmented-generation-for-medical-question-answering-with-llama-2-7b.

[12] A Liu and et al. Bias evaluation and mitigation in retrieval-augmented medical qa. *arXiv preprint arXiv:2503.15454*, 2024.

[13] A Roberts and et al. Influence of context size and model choice in rag systems. In *Findings of NAACL*, 2025.

[14] Z Wang and et al. Medgraphrag: Safe medical llm via graph-based retrieval-augmented generation. *arXiv preprint arXiv:2408.04187*, 2024.

[15] X Zhao and et al. Aipatient: Simulating patients with ehrs and agentic workflow. *arXiv preprint arXiv:2409.18924*, 2024.

A APPENDIX: Prompts Used in the Surgical Information Assistant

This appendix contains all the prompts used in the Surgical Information Assistant codebase. These prompts are used to guide the language models in performing various tasks such as question decomposition, information retrieval, answer synthesis, and follow-up question generation.

A.1 Question Decomposition Prompt

This prompt is used to break down complex surgical questions into simpler sub-questions:

```
You are an expert at breaking complex surgical questions into simpler ones. Break the
    following question into smaller sub-questions:

Question: {question}

Each sub-question should be independent and answerable on it's own without needing
    reference to other sub-questions. Think of at least 3 sub-questions but no more than
    7.

Think step-by-step and make sure to reason through how break the question into
    sub-questions.

Create new sub-questions in the following format but do NOT answer the question. Respond
    in the following format:

<think> Your reasoning here... </think>
<sub-question> The first sub-question... </sub-question>
<sub-question> The second sub-question... </sub-question>
...
<sub-question> The last sub-question... </sub-question>
```

A.2 Answer Generation from Context Prompt

This prompt is used to generate answers from retrieved context:

```
Based on the given question and context, generate an answer.
Question: {question}
Context: {context}

Think step-by-step and make sure to reason through how to generate an answer. ONLY rely
    on the given context to generate the answer.

Include snippets of the context that support your answer. Do NOT use any information
    outside of the given context to generate the answer.

Respond in the following format:

<think> Your reasoning here... </think>
<answer> The generated answer... </answer>
<snippet> First relevant snippet from the context... </snippet>
<snippet> Second relevant snippet from the context... </snippet>
...
<snippet> The last relevant snippet from the context </snippet>
```

A.3 Answer Synthesis Prompt

This prompt is used to synthesize answers from multiple sub-questions:

```
You are a reasoning engine. Given the following sub-question answers, determine whether
    they are enough to fully answer the original question. ONLY rely on the knowledge to
    determine whether the question can be answered.

If yes, then provide the answer. Make your answer detailed and structured with sections,
    as appropriate. Include as much relevant information as possible from the knowledge
    provided.

If you determine that you cannot answer the original question, then suggest what
    additional questions should be asked.

Original Question:
{original_question}

Knowledge:
{answers}

Think step-by-step to reason through you answer and consider the relevant information.
    Respond in the following format:
<think> Your reasoning here... </think>
<can_answer> yes OR no </can_answer>
<answer> The answer to the original question... </answer>
<new_questions> The first new sub-question... </new_questions>
<new_questions> The second new sub-question... </new_questions>
...
<new questions> The last new sub-question </new_questions>
```

A.4 Best Effort Answer Generation Prompt

This prompt is used when the system needs to generate a best-effort answer using Wikipedia:

```
You are a reasoning engine. Given the following original question and sub-question
    answers, formulate an answer to the best of your ability.

Original Question:
{original_question}

Knowledge:
{state["answers"]}
{state["wikipedia_results"]}

Think step-by-step to reason through you answer and consider the relevant information.
    Respond in the following format:
<think> Your reasoning here... </think>
<answer> The answer to the original question... </answer>
```

A.5 Follow-up Question Generation Prompt

This prompt is used to generate follow-up questions:

```
You are a reasoning engine. Given the following original question and final answer,
    generate 3 follow-up questions that help expand on the original question and the
    answer in a step-wise manner.

Original Question:
{original_question}

Final Answer:
{final_answer}
```

```
Think step-by-step to reason through your answer and consider the relevant information.
    Respond in the following format:
<think> Your reasoning here... </think>
<follow_up_questions> follow-up question here... </follow_up_questions>
<follow_up_questions> follow-up question here... </follow_up_questions>
<follow_up_questions> follow-up question here... </follow_up_questions>
```

A.6 Chain-of-Thought Generation Prompt

This prompt is used to generate detailed reasoning for answers:

```
You are a reasoning engine. Based on the following question and knowledge, provide a
    detailed, step-by-step reasoning to arrive at an answer. Include at least 3 steps,
    but more as needed.

Question:
{state["original_question"]}

Knowledge:
{state["answers"]}
{state["wikipedia_results"] if "wikipedia_results" in state else ""}

Provide your response in this format:

<think> Your reasoning here... </think>
<answer> The final answer here... </answer>
```

A.7 Surgery Topic Classification Prompt

This prompt is used to determine if a question is about surgery:

```
Determine if the following question requires access to specific medical documents
    related to detailed surgical information to be answered accurately.
Think step-by-step and reason through your answer. Respond in the following format:

<thinking> Your reasoning here... </thinking>
<answer> yes OR no </answer>

Here are some examples:

Question:
What is the primary purpose of the coronary artery bypass graft?
Response:
<thinking> The question is asking about a coronary bypass graft which is related to
    surgery. So yes, this question is about surgery. </thinking>
<answer> yes </answer>

Question:
Is machine learning useful for solving complex medical problems?
Response:
<thinking> The question is asking about machine learning as a tool for solving complex
    medical problems. While the question does specify medical problems, it does not
    refer to surgery or topics related to surgery. </thinking>
<answer> no </answer>

Question:
Can you elaborate on your previous response about suturing or rephrase it?
Response:
```

```
<thinking> The question is asking for an explanation of suturing that was given earlier
    in the conversation. While the question does mention suturing which is related to
    surgery, it is not asking about suturing specifically but rather requesting to
    explain a previous response. </thinking>
<answer> no </answer>

Question:
What are some core differences between robotic and laparasopic inguinal hernia?
Response:
<thinking> The question is asking about the differences between robotic and laparosopic
    inguinal hernia. Both robotic and laparosopic surgery involve surgery on the
    inguinal hernia. This is clearly a question about surgery and requires investigation
    into documentation to answer. </thinking>
<answer> yes </answer>

Question:
{question}
Response:
```

A.8 Context-Aware Input Update Prompt

This prompt is used to update user input with context from chat history:

```
Given the following chat history and a new user input, update the user input to include
    any relevant context from the chat history. Only add necessary context, and keep the
    updated input concise.
Think step-by-step and reason through your answer. Respond in the following format:

<thinking> Your reasoning here... </thinking>
<updated_input> Updated user input here... </updated_input>

Here are some examples:

CHAT HISTORY:
User: What are the main types of sutures used in surgery?
Assistant: The main types of sutures used in surgery are absorbable and non-absorbable
    sutures. Absorbable sutures dissolve over time, while non-absorbable sutures need to
    be removed. Common absorbable sutures include catgut and synthetic materials like
    polyglycolic acid. Non-absorbable sutures are often made of materials like nylon or
    silk.
User: How long do absorbable sutures typically last?
Assistant: The duration of absorbable sutures varies depending on the material used.
    Generally, they can last anywhere from a few days to several months. For example,
    fast-absorbing sutures may dissolve within 7-10 days, while slower-absorbing sutures
    can take 60-90 days or even longer.
NEW USER INPUT:
What about non-absorbable ones?
RESPONSE:
<thinking> Considering the previous discussion the user is referring to non-absorbable
    sutures. Specifically, they are asking about the lifespan or removal time-frame for
    non-absorbable sutures. I'll restate the question to include this context from the
    chat history. </thinking>
<updated_input> What is the typical lifespan or removal timeframe for non-absorbable
    sutures used in surgery? </updated_input>



CHAT HISTORY:
User: What is a laparotomy?
```

```
Assistant: A laparotomy is a surgical procedure that involves making a large incision in
    the abdomen to gain access to the abdominal cavity. It allows surgeons to examine
    the abdominal organs, diagnose problems, and perform necessary treatments.
    Laparotomies are used for various purposes, including exploratory surgery, treatment
    of abdominal trauma, and removal of tumors.
User: Are there any alternatives to this procedure?
Assistant: Yes, there are alternatives to laparotomy, particularly minimally invasive
    techniques. The main alternative is laparoscopy, also known as keyhole surgery. In
    laparoscopy, several small incisions are made instead of one large incision. A
    camera and specialized instruments are inserted through these small incisions to
    perform the surgery. This technique often results in less pain, faster recovery, and
    smaller scars compared to traditional laparotomy.

NEW USER INPUT:
What are the risks?

RESPONSE:
<thinking> The user is asking about risks, but it's not clear whether they're referring
    to laparotomy or laparoscopy, both of which were discussed in the previous messages.
    Since laparotomy was the original topic and laparoscopy was introduced as an
    alternative, it would be helpful to ask about the risks of both procedures for a
    comprehensive answer. </thinking>
<updated_input> What are the risks associated with both laparotomy and laparoscopy
    procedures? </updated_input>



CHAT HISTORY:
User: What are the potential complications of laparoscopic surgery?
Assistant: Laparoscopic surgery can have several potential complications, including:
1. Infection: Laparoscopic surgery can lead to infections, especially in patients with
    infections that can be spread through the abdominal cavity.
2. Stool bleeding: Laparoscopic surgery can lead to stool bleeding, especially if the
    surgical instrument is inserted directly into the abdominal cavity.
3. Pain: Laparoscopic surgery can be painful, especially for patients with pre-existing
    conditions or those who have had previous laparoscopic surgeries.

NEW USER INPUT:
Can you think of any more?

RESPONSE:
<thinking> The user is asking if there are any more complications, but it's important to
    note that the original question was about laparoscopic surgery, not about potential
    complications. I'll restate the question to clarify that the original topic was
    laparoscopic surgery. </thinking>
<updated_input> Are there any more potential complications associated with laparoscopic
    surgery besides infection, stool bleeding, and pain? </updated_input>



CHAT HISTORY:
{formatted_history}

New User Input:
{user_input}

RESPONSE:
```

A.9 Answer Evaluation Prompt

This prompt is used to evaluate the correctness of generated answers:

```
You are a medical reasoning engine that compares two answers to a given question to
    determine whether the answers are the same. Here is the question and the two answers:

Question:
{state["original_question"]}

Answer 1:
{known_answer}

Answer 2:
{state['final_answer']}

Think step-by-step and provide a detailed reasoning process that compares the two
    answers given the context of the question. Include at least 3 steps in your
    reasoning, but more as needed.

Respond in this format:

<thinking> Your reasoning here... </thinking>
<answer> yes OR no </answer>
```

A.10 Best Effort Answer Generation with Wikipedia Prompt

This prompt is used when the system needs to generate a best-effort answer using Wikipedia:

```
You are a reasoning engine. Given the following original question and sub-question
    answers, formulate an answer to the best of your ability.

Original Question:
{original_question}

Knowledge:
{answers}
{wikipedia_results}

Think step-by-step to reason through you answer and consider the relevant information.
    Respond in the following format:
<think> Your reasoning here... </think>
<answer> The answer to the original question... </answer>
```

A.11 Simple Question Handling Prompt

This prompt is used to handle simple questions that don't require access to surgical documents:

```
You are a helpful assistant that answers questions about general topics. Please respond
    to the following question:

{question}

If the question is related to surgery or medical procedures, please provide a general
    answer based on your knowledge, but note that for detailed surgical information, it
    would be better to consult medical documents.

Think step-by-step and reason through your answer. Provide a clear, concise, and
    accurate response.
```

A.12 Wikipedia Search Prompt

This prompt is used to generate search queries for Wikipedia:

```
Given the following question, generate 3 search queries that would help find relevant
    information on Wikipedia. The queries should be specific and focused on the key
    concepts in the question.

Question: {question}

Respond in the following format:
<query>first search query</query>
<query>second search query</query>
<query>third search query</query>
```

A.13 Wikipedia Context Synthesis Prompt

This prompt is used to synthesize information from Wikipedia contexts:

```
You are a reasoning engine. Given the following question and Wikipedia contexts,
    synthesize the information to provide a comprehensive answer.

Question: {question}

Wikipedia Contexts:
{contexts}

Think step-by-step to reason through your answer and consider the relevant information
    from the contexts. Respond in the following format:
<think> Your reasoning here... </think>
<answer> The synthesized answer... </answer>
```

A.14 ReAct Agent System Prompt

This prompt is used for the ReAct agent in the evaluation scripts:

```
You are a medical assistant specializing in surgical information. Your goal is to
    provide accurate and helpful information about surgical procedures, techniques, and
    related medical knowledge.

When answering questions:
1. Use the DocumentSearch tool to find relevant information in the surgical database
2. Analyze the retrieved information carefully
3. Provide comprehensive and accurate answers based on the retrieved content
4. If the information is not available in the search results, acknowledge the limitations
5. Think step-by-step to reason through complex questions
6. Cite specific parts of the retrieved documents when appropriate
7. Focus on providing factual medical information rather than opinions

Remember accuracy is crucial. Provide all reasoning and the final answer.
```

A.15 Question-Answer Pair Generation Prompt

This prompt is used to generate question-answer pairs from text chunks:

```
You are a medical reasoning engine. Given a medical or medically-related text passage,
    you must generate question-answer pairs. The questions MUST be answerable by the
    information in the passge but understandable on its own without needing direct
    reference to the passage (ie. "what is this passage about?" is NOT a valid question).
```

```
The questions should have enough information to be able to retrieve the relevant
    passages in the future to help answer the question.

Think step-by-step and reason through the content of the passage to hypothesize
    potential question-answer pairs and then respond. Think of at least 8
    question-answer pairs but more, if possible.

Respond in this format:
<think> You reasoning here... </think>
<qa-pair> [question here...] | [answer here...] </qa-pair>
<qa-pair> [question here...] | [answer here...] </qa-pair>
```

B Prompt Design Principles

The prompts used in the Surgical Information Assistant follow several key design principles:

B.1 Step-by-Step Reasoning

All prompts encourage the language model to think step-by-step and provide detailed reasoning before arriving at an answer. This approach, often referred to as chain-of-thought prompting, has been shown to improve the accuracy and reliability of language model outputs, especially for complex tasks.

B.2 Structured Output Format

The prompts use a consistent structured output format with XML-like tags (e.g., `<think>`, `<answer>`, `<sub-question>`) to clearly separate different components of the response. This structured format makes it easier to parse and extract specific information from the language model's output.

B.3 Few-Shot Examples

Several prompts include few-shot examples to demonstrate the expected reasoning process and output format. These examples help guide the language model to produce responses that follow the desired pattern and meet the quality standards.

B.4 Task-Specific Instructions

Each prompt contains clear, task-specific instructions that define the role of the language model (e.g., "You are an expert at breaking complex surgical questions into simpler ones") and the specific task it needs to perform. This helps focus the model's attention on the relevant aspects of the task.

B.5 Context Awareness

The prompts are designed to make the language model aware of the context in which it operates, including the original question, retrieved information, and previous interactions. This context awareness is crucial for generating coherent and relevant responses.

C Prompt Usage in the System Architecture

The prompts are used at different stages of the Surgical Information Assistant's workflow:

1. **Question Processing**: The system first determines if a question is about surgery using the Surgery Topic Classification Prompt. If it is, the question is processed through the agent-based workflow. If not, it is handled using the Simple Question Handling Prompt.

2. **Question Decomposition**: Complex surgical questions are broken down into simpler sub-questions using the Question Decomposition Prompt.

3. **Information Retrieval**: For each sub-question, relevant information is retrieved from the surgical document database.

4. **Answer Generation**: The Answer Generation from Context Prompt is used to generate answers for each sub-question based on the retrieved context.

5. **Answer Synthesis**: The Answer Synthesis Prompt is used to combine the answers to the sub-questions into a comprehensive answer to the original question.

6. **Best Effort Answer**: If the system cannot fully answer the question with the available documents, it uses the Best Effort Answer Generation Prompt to generate a best-effort answer, potentially using information from Wikipedia.

7. **Follow-up Question Generation**: After providing an answer, the system generates follow-up questions using the Follow-up Question Generation Prompt to help users explore the topic further.

8. **Chain-of-Thought Generation**: The Chain-of-Thought Generation Prompt is used to provide detailed reasoning for the final answer.

This structured approach to prompt design and usage enables the Surgical Information Assistant to provide accurate, comprehensive, and contextually relevant answers to a wide range of surgical questions.