

# Medical Assistant

PGP-AIML-BA-UTA-Jun25-D

Date 12/11/2025

*Yashpal Singh*

## Contents / Agenda

- Question Answering using LLM
- Question Answering using LLM with Prompt Engineering
- Data Preparation for RAG
- Question Answering using RAG
- Output Evaluation

# Executive Summary:

## AI-Powered Medical Knowledge Retrieval System

### Business Problem & Project Context :

The modern healthcare landscape faces a critical challenge of Information Overload, where the volume of medical data exceeds the manual synthesis capacity of healthcare professionals .

**Key pain points include:**

- **Information Fragmentation:** Knowledge is scattered across disparate manuals and journals.
- **Time-Critical Decisions:** Emergency and critical care settings require rapid processing of complex information.
- **Variability in Care:** Reliance on memory leads to inconsistent diagnosis and treatment protocols.

**Primary Objectives** The project aims to develop an AI prototype utilizing **Retrieval-Augmented Generation (RAG)** to provide a "**single source of truth**" from verified medical literature. The core goal is to enhance diagnostic accuracy and standardize care through evidence-based, traceable AI responses..

**Key Performance Insights (Evaluation Results) :** Using an LLM-as-a-Judge approach (Mistral-7B), the system was evaluated across three developmental stages:

- **Baseline LLM (No RAG):** Responses were found to be overly generic and clinically shallow. A significant issue was abrupt truncation due to low token limits (128 tokens), resulting in incomplete medical advice.
- **LLM with Prompt Engineering:** The addition of a "Medical Consultant" system prompt improved structural integrity and clinical tone. However, without a dedicated knowledge base, the model still missed core clinical specifics (e.g., failed to answer if appendicitis can be cured with medicine).
- **RAG Implementation:** This stage where various parameters like **token size =512** and **temperature =0** are adjusted showed the highest **Groundedness** and **Relevance (rating 5/5 for sepsis protocols)**. By retrieving from the **medical\_diagnosis\_manual.pdf**, the system provided specific, evidence-based details like specific antibiotic regimens (e.g., Vancomycin, Ceftazidime) that the base LLM could not produce.

# Executive Summary:

## Actionable insights and recommendations

### Technical Recommendations:

- ✓ **Resolve Output Truncation:** Across all testing phases, the model frequently cut off mid-sentence. **Recommendation:** Increase max\_tokens to at least 512 for complex medical protocols to ensure complete, actionable clinical guidance.
- ✓ **Optimize Hyperparameters for Stability:** A Temperature of **0** provided high groundedness and deterministic accuracy for protocols. While a Temperature of **0.5** offered more descriptive summaries, it sometimes lacked the precision required for high-stakes medical scenarios
- ✓ **Enforce Structural Formatting:** Despite prompt instructions, the RAG model occasionally failed to use required headers like `###Answer`. **Recommendation:** Refine the "Few-Shot" examples in the prompt template to strictly enforce output schema

### Strategic & Clinical Recommendations:

- ✓ **Expand the Knowledge Base:** The RAG system's success is directly tied to the quality of retrieved chunks. The current prototype relies on a single manual. **Recommendation:** Integrate diverse data sources (WHO standards, internal hospital guidelines) to reduce "I don't know" responses and broaden the tool's utility.
- ✓ **Human-in-the-Loop Validation:** While the "LLM-as-a-Judge" approach is efficient, it showed the model suggesting slightly excessive measures (e.g., C-collar for a simple leg fracture). **Recommendation:** Implement a phase of expert medical review to fine-tune the "Relevance" and "Safety" of AI recommendations before clinical deployment.
- ✓ **Standardize "Evidence-Based" Responses:** Use the RAG system to automate the generation of standardized care pathways, which addresses the "Variability Care" pain point identified in the problem statement.

# Business Problem Overview- Medical Assistant

The modern healthcare landscape is defined by an exponential growth in medical knowledge, leading to a critical challenge: **Information Overload** and the resultant strain on timely, accurate clinical decision-making.

## **The Challenge Information Overload in Healthcare:**

Healthcare professionals (HCPs) operate in a highly demanding, time-sensitive environment where the volume of medical data—from new research to constantly updated clinical guidelines—far exceeds the capacity for manual retrieval and synthesis.

Key Pain Point	Description	Impact on Patient Care
<b>Information Fragmentation</b>	Medical knowledge is scattered across various sources (manuals, journals, internal documents).	Medical knowledge is scattered across various sources (manuals, journals, internal documents).
<b>Time Critical Decision</b>	Especially in emergency/critical care settings, minutes count, but complex information requires quick processing	Especially in emergency/critical care settings, minutes count, but complex information requires quick processing
<b>Variability Care</b>	Reliance on individual memory and disparate sources can lead to inconsistencies in diagnosis and treatment protocols	Lack of standardized, evidence-based care across different practitioners or facilities

# Business Problem Overview- Medical Assistant

## Project Objectives and Value proposition :

The primary objective is **to streamline access to complex medical information** to enhance diagnostic accuracy, treatment planning, and overall operational efficiency.

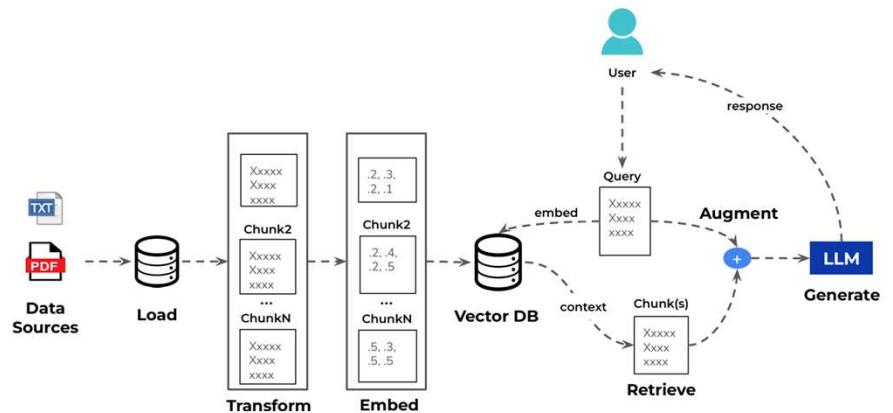
Component	Goal	Value to Healthcare
RAG System Development	<b>Develop</b> a functional <b>AI prototype</b> using established medical knowledge as the single source of truth.	<b>Reliable and Trusted Information:</b> Ensure answers are traceable and sourced from verified medical literature
Decision Support	<b>Apply</b> AI to quickly generate evidence-based answers to complex clinical questions (e.g., sepsis protocols, surgical procedures).	<b>Improved Efficiency:</b> Reduce time spent on information retrieval, allowing more focus on patient care.
Impact and Evolution	<b>Analyze and Evaluate</b> the solution's potential to standardize care and improve patient outcomes through faster, more informed decisions.	<b>Standardized Care:</b> Promote consistent adherence to best practice guidelines across the organization

This project directly addresses the need for **efficiency, accuracy, and standardization** in a data-rich clinical setting, moving healthcare centers toward a more informed, **data-driven model of patient care**

# Solution Approach-RAG

To bridge the gap between vast medical knowledge and point-of-care decisions, I propose developing an **AI-Powered Medical Knowledge Retrieval and Decision Support System utilizing a Retrieval-Augmented Generation (RAG) architecture by following below steps which has depicted in this diagram**

- 1. Upload Data Sources :** Ingest medical reference documents ( [medical\\_diagnosis\\_manual.pdf](#) ) from the designated directory.
- 2. Document Chunking :** Split the documents into manageable, semantically coherent chunks using the [RecursiveCharacterTextSplitter](#) library.
- 3. Embedding Generation :** Convert each text chunk into vector embeddings using [SentenceTransformerEmbeddings](#) with the embedding model [thenlper/gte-large](#).
- 4. Vector Storage :** Store the generated embeddings in a vector database (e.g., [MedicalDB](#)) using [Chroma](#) for efficient similarity search.
- 5. Context Retrieval :** Utilize a retriever component to fetch the most relevant chunks from the vector database based on the user query
- 6. Query Augmentation and generation** Augment the user query with the retrieved contextual chunks and pass the enriched prompt to the LLM([Mistral-7B-Instruct-v0.2-CKUF](#)) to generate final response



## Evaluation Mechanism :

- ✓ The *LLM-as-a-Judge* approach is employed to evaluate the quality of the RAG system across two key dimensions: **retrieval quality** and **generation quality**.
- ✓ The same Mistral model is used for both response generation and evaluation;
- ✓ thus, the LLM assesses its own outputs with respect to **groundedness** (alignment with the retrieved context) and **relevance** (alignment with the user query).

# Question Answering using LLM

## LLM Details :

**model\_name\_or\_path** = "TheBloke/Mistral-7B-Instruct-v0.2-GGUF"

**model\_basename** = "mistral-7b-instruct-v0.2.Q6\_K.gguf"

**Parameters :** *max\_tokens=128,temperature=0,top\_p=0.95,top\_k=50*

## **What treatment options are available for managing hypertension?**

### Response received :

Hypertension, or high blood pressure, is a common condition that can increase the risk of various health problems such as heart disease, stroke, and kidney damage. The good news is that there are several effective treatment options available to help manage hypertension and reduce the risk of complications. Here are some of the most commonly used treatments:

1. Lifestyle modifications: Making lifestyle changes is often the first line of defense against hypertension. This may include eating a healthy diet rich in fruits, vegetables, whole grains, and lean proteins; limiting sodium intake; getting regular physical activity

### Comment :

**General Response :** I can observe the LLM has given very general response and very common symptom for Hypertension , like high blood pressure and very generic treatment like Lifestyle modifications instead some appropriate prescribed medicines

# Question Answering using LLM

**Parameters :** *max\_tokens=128,temperature=0,top\_p=0.95,top\_k=50*

## Question 1 What is the protocol for managing sepsis in a critical care unit?

Sepsis is a life-threatening condition that can arise from an infection, and it requires prompt recognition and aggressive management in a critical care unit. The following are general steps for managing sepsis in a critical care unit:

1. Early recognition: Recognize the signs and symptoms of sepsis early and initiate treatment as soon as possible. Sepsis can present with various clinical features, including fever or hypothermia, tachycardia or bradycardia, altered mental status, respiratory distress, and lactic acidosis.
2. ABCs

## Comment :

**Generic and Irrelevant Response:** The model's answer was overly generic, and Point 2 was entirely irrelevant (e.g., "ABC"). While the question specifically asked for a management protocol, the LLM instead provided information on sepsis and its recognition, failing to address the intended scope of the query.

# Question Answering using LLM

**Parameters :** *max\_tokens=128,temperature=0,top\_p=0.95,top\_k=50*

**Question 2 What are the common symptoms for appendicitis, and can it be cured via medicine? If not, what surgical procedure should be followed to treat it?**

## Response Received

Appendicitis is a medical condition characterized by inflammation of the appendix, a small tube-shaped organ located in the lower right side of the abdomen. The symptoms of appendicitis can vary from person to person, but some common signs include:

1. Abdominal pain: The pain is typically located in the lower right quadrant of the abdomen and may start as a mild discomfort that gradually worsens over time. The pain may be constant or intermittent and may be aggravated by movement, deep breathing, or coughing.
2. Loss of appetite.

## **Comment :**

**Generic Response:** The model provide general information for symptoms for appendicitis but not recommending any actual medicines

**Abrupt truncation :** Model suddenly got truncated **at point number 2 Loss of appetite** due to the token limit

# Question Answering using LLM

**Parameters :** *max\_tokens=128,temperature=0,top\_p=0.95,top\_k=50*

**Question 3. What are the effective treatments or solutions for addressing sudden patchy hair loss, commonly seen as localized bald spots on the scalp, and what could be the possible causes behind it?**

## Response Received

Sudden patchy hair loss, also known as alopecia areata, is a common autoimmune disorder that affects the hair follicles. It can result in round or oval bald patches on the scalp, but it can also occur on other parts of the body such as the beard area, eyebrows, and eyelashes.

The exact cause of alopecia areata is not known, but it's believed to be related to a problem with the immune system. Some possible triggers for this condition include stress, genetics, viral infections, and certain medications.

## Comment :

**Generic and Incomplete Response:** The model failed to provide information on effective treatment options for patchy hair loss. Additionally, it did not adequately address the core question regarding potential causes of the condition, instead offering a generic statement that the exact cause of alopecia areata is unknown, without further explanation or contextual detail

# Question Answering using LLM

**Parameters:** `max_tokens=128,temperature=0,top_p=0.95,top_k=50`

**Question 4.** What treatments are recommended for a person who has sustained a physical injury to brain tissue, resulting in temporary or permanent impairment of brain function?

## **Response received**

A person who has sustained a physical injury to the brain, also known as a traumatic brain injury (TBI), may require various treatments depending on the severity and location of the injury. Here are some common treatments recommended for individuals with TBIs:

1. Emergency care: In case of a severe TBI, the first priority is to provide emergency care to ensure the person's airway is clear, breathing is stable, and circulation is adequate. This may involve intubation, oxygen therapy, and intravenous fluids.
2. Surgery: Depending on the location and

## **Comment:**

1. **Generic and Incomplete Response:** The model failed to provide information on effective treatment options for sustained a physical injury to brain tissue, instead offering a generic statement that A person who has sustained a physical injury to the brain, also known as a traumatic brain injury (TBI), without further explanation or contextual detail
2. **Abrupt truncation :** Model suddenly got truncated at point number 2 Surgery: Depending on the location and due to the token limit

# Question Answering using LLM

**Parameters :** max\_tokens=128,temperature=0,top\_p=0.95,top\_k=50

**Question 5. What are the necessary precautions and treatment steps for a person who has fractured their leg during a hiking trip, and what should be considered for their care and recovery?**

## **Response Received**

First and foremost, if you suspect that someone has fractured their leg while hiking, it's essential to ensure their safety and prevent further injury. Here are some necessary precautions:

1. Keep the person calm and still: Encourage them to remain as still as possible to minimize pain and prevent worsening the injury.
2. Assess the situation: Check for any signs of shock, such as pale skin, rapid heartbeat, or shallow breathing. If you notice these symptoms, seek medical help immediately.
3. Immobilize the leg: Use a splint, sl

## **Comments :**

1. **Generic and Incomplete Response:** The model failed to provide information on effective treatment options for a person who has fractured their leg during a hiking trip, and what should be considered for their care and recovery, instead offering a generic statement that First and foremost, if you suspect that someone has fractured their leg while hiking, it's essential to ensure their safety and prevent further injury
2. **Abrupt truncation :** Model suddenly got truncated at point number 3 Immobilize the leg: Use a splint, sl

# Question Answering using LLM with Prompt Engineering



## **System Prompt :**

You are a highly experienced Medical Consultant and Critical Care Specialist.  
Your task is to provide detailed, evidence-based medical information based on current clinical guidelines (such as the Surviving Sepsis Campaign or WHO standards).

For each question provided:

1. Use clear headings for different sections of the answer.
2. Maintain a professional, clinical, yet accessible tone.
3. Distinguish clearly between diagnostic signs, immediate interventions, and long-term recovery.
4. Include a standard medical disclaimer at the end.

## **Question 1.**

**What is the protocol for managing sepsis in a critical care unit?**

## **Response Received :**

\*\*Section 1: Definition and Recognition of Sepsis\*\*

Sepsis is a life-threatening condition caused by a dysregulated host response to infection. It can lead to tissue damage, multiple organ failure, and even death if not recognized and treated promptly. The definition of sepsis has evolved over the years, with the most recent update being from the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) in 2016.

\*\*Section 2: Early Recognition and Diagnosis\*\*

## **Comments :**

- **Structural Integrity :** This time LLM structured the answer in various sections based on System prompt
- **Clinical Accuracy and Alignment :** It adhered well to the instruction to use clear headings, improving scan ability for complex protocols like ICU management.
- **Conciseness vs. Depth :** While the recognition phase is accurate, **the model stopped due to token limits**

# Question Answering using LLM with Prompt Engineering

## Question 2

**What are the common symptoms for appendicitis, and can it be cured via medicine? If not, what surgical procedure should be followed to treat it??**

### **Response Received:**

\*\*Symptoms of Appendicitis:\*\*

Appendicitis is a common inflammatory condition characterized by the swelling and eventual rupture of the appendix. The following are the most common symptoms:

1. \*\*Perumbilical or right lower quadrant abdominal pain:\*\* This is often described as a dull ache that may progress to sharp, localized pain in the right lower quadrant of the abdomen.
2. \*\*Anorexia and nausea:\*\* Loss of appetite and feeling sick to your stomach are common symptoms. Vomiting may occur

### **Comments :**

- **Clinical Detail & Accuracy:** The model correctly identified the **classic migrating pain** (perumbilical to right lower quadrant abdominal pain) and associated gastrointestinal symptoms (**anorexia, nausea**), which are hallmark diagnostic indicators.
- **Abrupt Truncation:** The response cut off mid-sentence ("**Vomiting may occur**"), due to token
- **Missed Core Question:** It failed to address the most critical part of your request: whether it can be cured with medicine (antibiotics) and the specifics of the surgical procedure.

# Question Answering using LLM with Prompt Engineering

## Question 3

**What are the effective treatments or solutions for addressing sudden patchy hair loss, commonly seen as localized bald spots on the scalp, and what could be the possible causes behind it?**

### Response Received :

\*\*Section 1: Possible Causes of Sudden Patchy Hair Loss\*\*

Sudden patchy hair loss, also known as alopecia areata, is an autoimmune disorder that results in the sudden onset of round or oval bald patches on the scalp. The exact cause of this condition is unknown, but several factors have been implicated:

1. \*\*Genetics:\*\* Alopecia areata has a strong genetic component, with up to 20% of cases being hereditary.
2. \*\*Immune System Dysfunction:\*\*

### Comments :

- **Clinical Detail & Accuracy:** The model correctly identified Alopecia Areata as the primary cause and accurately defined its autoimmune nature
- **Logical Categorization:** It successfully distinguished between genetic predisposition and immune system dysfunction, which are the two most critical pillars in understanding this condition's etiology.
- **Abrupt Truncation** Section: The model cut off before reaching the "Effective Treatments" part of the question, **due to the token limit**

# Question Answering using LLM with Prompt Engineering

## Question 4.

**What treatments are recommended for a person who has sustained a physical injury to brain tissue, resulting in temporary or permanent impairment of brain function?**

### Response Received

\*\*Heading 1: Initial Assessment and Stabilization\*\*

In case of a suspected traumatic brain injury (TBI), the first priority is to ensure the safety and stability of the patient. This includes securing the airway, maintaining adequate ventilation, and controlling any bleeding. The following interventions are recommended based on current clinical guidelines:

1. \*\*Airway Management:\*\* If there's a risk of airway obstruction or aspiration, secure the airway using endotracheal intubation. Monitor the patient's level of consciousness (LOC) and oxygen

### Comments :

- **Structural Integrity :** This time LLM structured the answer in various sections based on System prompt
- **Clinical Accuracy of Metrics:** It correctly linked the level of consciousness (LOC) to airway management, a critical decision point in neurocritical care (e.g., "GCS less than 8, intubate").
- **Abrupt Truncation :** The response ends mid-sentence ("oxygen...") due to the token limit

# Question Answering using LLM with Prompt Engineering

## Question 5

**What are the necessary precautions and treatment steps for a person who has fractured their leg during a hiking trip, and what should be considered for their care and recovery?**

### Response Received

\*\*Necessary Precautions:\*\*

1. \*\*Immediate Assessment:\*\* Evaluate the severity of the injury by assessing the level of pain, swelling, deformity, open wounds, and signs of neurovascular compromise such as pulselessness, pallor, or paralysis. If there are signs of severe injury, immobilize the leg using a splint or a C-collar and transport the patient to the nearest emergency department as soon as possible.
2. \*\*Maintain Immobility:\*\* Avoid moving the fractured leg excessively to prevent

### Comments :

- **Structural Response :** The model correctly structure it in various sections(\*\*Necessary Precautions:\*\* , Immediate Assessment:--) based on the instruction provided
- **Medical Terminology :** Interestingly, the model suggests using a C-collar; while generally used **for spinal trauma**, it implies the model is considering the "mechanism of injury" (a fall while hiking), though it is slightly excessive for a simple isolated leg fracture.
- **Abrupt Truncation** The response **cuts off at the exact moment** it begins to explain how to prevent further injury ("prevent...") due to the token limit

## Data Preparation for RAG

- **Dataset used : medical\_diagnosis\_manual.pdf**
- **Parameters used to split the data:**
  - ✓ `chunk_overlap=20`
  - ✓ `chunk_size=512`
- **Name of the embedding model used: '`thenlper/gte-large`'**
- **Parameters used to define the RAG**
  - ✓ `k=3`
  - ✓ `max_tokens =256`
  - ✓ `Temperature=0`
  - ✓ `top_p=0.95`
  - ✓ `top_k=50`

# Question Answering using RAG

## **System Prompt :**

You are an assistant whose work is to review the report and provide the appropriate answers from the context.  
User input will have the context required by you to answer user questions.

This context will begin with the token: ###Context.

The context contains references to specific portions of a document relevant to the user query.

User questions will begin with the token: ###Question.

Please answer only using the context provided in the input. Do not mention anything about the context in your final answer.

If the answer is not found in the context, respond "I don't know".

## **User Prompt template:**

###Context

Here are some documents that are relevant to the question mentioned below.

{context}

###Question

{question}

# Question Answering using RAG

## **Question 1.**

What is the protocol for managing sepsis in a critical care unit?

## **Response Received :**

Based on the context provided, the protocol for managing sepsis in a critical care unit includes:

1. Administering parenteral antibiotics after taking specimens for Gram stain and culture.
2. Starting very prompt empiric therapy as soon as sepsis is suspected.
3. Selecting an antibiotic regimen based on the suspected source, clinical setting, knowledge or suspicion of causative organisms and sensitivity patterns common to that specific inpatient unit, and previous culture results.
4. Adding vancomycin if resistant staphylococci or enterococci are suspected.
5. Including a drug effective against anaerobes if there is an abdominal source.
6. Changing the antibiotic regimen based on culture and sensitivity results.
7. Continuing antibiotics for at least 5 days after shock resolves and evidence of infection subsides.
8. Draining abscesses and excising necrotic tissues to eliminate septic foci.
9. Normalizing blood glucose levels between 80 to 110 mg/dL (4.4 to 6.

## **Comments :**

- **Lack of Structural Integrity :** Built RAG model mostly followed the System prompt and user template and listed 9 protocols for managing sepsis in a critical care unit, but **expectation was Answer will be starting with ###Answer** and that not happen
- **Clinical Accuracy and Alignment :** It adhered well and clearly provided appropriate information from the dataset medical manual for example (“*Selecting an antibiotic regimen based on the suspected source, clinical setting, knowledge or suspicion of causative organisms and sensitivity patterns common to that specific inpatient unit, and previous culture results.*”)

# Question Answering using RAG

## Question 2

What are the common symptoms for appendicitis, and can it be cured via medicine? If not, what surgical procedure should be followed to treat it??

### Response Received:

###Answer

The common symptoms for appendicitis include abdominal pain, anorexia, and abdominal tenderness. Appendicitis cannot be cured via medicine alone; surgery, specifically a surgical removal of the appendix (**appendectomy**), is required for treatment.

### Comments :

- **Structural Integrity :** Built RAG model mostly followed the System prompt and user template for above question and **Answer started with ###Answer**
- **Clinical Detail & Accuracy:** The model correctly identified the **classic migrating pain** (perumbilical to right lower quadrant abdominal pain ) and associated gastrointestinal symptoms (**anorexia, nausea**), which are hallmark diagnostic indicators.
- **Abrupt Truncation:** The response was cut off which may indicate a token limit issue or an artifact of the temperature setting (Temp = 0).
- **Core Question :**what surgical procedure should be followed to treat is answered in very brief ("Appendicitis cannot be cured via medicine alone; surgery, specifically a surgical removal of the appendix (**appendectomy**), is required for treatment")

# Question Answering using RAG

## Question 3

What are the effective treatments or solutions for addressing sudden patchy hair loss, commonly seen as localized bald spots on the scalp, and what could be the possible causes behind it?

### Response Received :

Based on the context provided, the condition being described is Alopecia Areata. The effective treatments for this condition include topical, intralesional, or systemic corticosteroids, topical minoxidil, topical anthralin, topical immunotherapy (diphencyprone or squaric acid dibutylester), or psoralen plus ultraviolet A (PUVA). The possible cause behind sudden patchy hair loss in this context is an autoimmune disorder.

### Comments :

- **Lack of Structural Integrity :** The built RAG model generally adhered to the system prompt and user template for Question 3; however, it failed to meet the expected response structure. Specifically, the answer did not begin with the **required ### Answer heading**
- **Clinical Detail & Accuracy:** The model correctly identified Alopecia Areata as the primary cause and accurately defined its autoimmune nature
- **Logical Categorization:** It successfully distinguished between genetic predisposition and immune system dysfunction, which are the two most critical pillars in understanding this condition's etiology.
- **Abrupt Truncation** Section: The model cut off before reaching the "Effective Treatments" part of the question, **due to the token limit** or an artifact of the **temperature setting (Temp = 0)**.

# Question Answering using RAG

## **Question 4.**

What treatments are recommended for a person who has sustained a physical injury to brain tissue, resulting in temporary or permanent impairment of brain function?

## **Response Received**

Based on the context provided, the following treatments are recommended for a person who has sustained a physical injury to brain tissue, resulting in temporary or permanent impairment of brain function:

1. Initial treatment: Ensuring a reliable airway and maintaining adequate ventilation, oxygenation, and blood pressure. Surgery may be needed in patients with more severe injury to place monitors to track and treat intracranial pressure, decompress the brain if intracranial pressure is increased, or remove intracranial hematomas.
2. Supportive care: Preventing systemic complications due to

## **Comments :**

- **Absence of Structural Integrity :** The built RAG model mostly obeyed to the system prompt and user template for **Question 4;** however, it miscarried to meet the expected response structure. Specifically, the answer did not **begin with the ###Answer Heading**
- **Clinical Accuracy :** It correctly linked the level of consciousness (LOC) to airway management, a critical decision point in neurocritical care (e.g., "GCS less than 8, intubate").
- **Abrupt Truncation :** The response ends mid-sentence ("oxygen...") due to the token limit

# Question Answering using RAG

## Question 5

What are the necessary precautions and treatment steps for a person who has fractured their leg during a hiking trip, and what should be considered for their care and recovery?

## Response Received

Based on the context provided, here is the answer:

The person with a fractured leg should elevate the injured limb above heart level for the first 2 days to minimize swelling. After 48 hours, they can apply warmth using a heating pad for 15 to 20 minutes to relieve pain and speed healing. Immobilization is necessary to prevent further injury and facilitate healing. Joints proximal and distal to the injury should be immobilized using either a cast or a splint. A cast is usually used for fractures that require weeks of immobilization, while a splint can be used for injuries that only need immobilization for several days. The person should keep the cast dry, never put an object inside it, inspect the edges and skin around the cast daily, apply lotion to any red or sore areas, pad any rough edges with soft material, and seek medical care if they notice an odor from within the cast or a fever indicating infection. Good hygiene is important. If severe swelling is likely, the cast may be cut open from end to end medially and laterally (bivalved). Prolonged immobilization of a joint can cause stiff

## Comments :

- **Lack of Structural Integrity :** The built RAG model generally adhered to the system prompt and user template for **Question 5**; however, it failed to meet the expected response structure. Specifically, the answer did not begin with the **###Answer**
- **Medical Terminology :** Interestingly, the model suggests Apply warmth (using a heating pad) for 15 to 20 minutes after the first 48 hours to relieve pain and speed healing
- **Abrupt Truncation** The response **cuts off at the exact moment** Prolonged immobilization of a joint can cause stiff --**due to the token limit**

# Question Answering using RAG with Fine tuning

**Parameters :** *max\_tokens=256,temperature=0.5,top\_p=0.95,top\_k=50*

## Question 1.

**What is the protocol for managing sepsis in a critical care unit?**

### **Response Received :**

The context provides information on managing sepsis in a critical care unit, including recommended antibiotic regimens and the importance of draining abscesses and eliminating septic foci. The text suggests that prompt empiric therapy is essential and may be lifesaving. Antibiotics should be started immediately after suspecting sepsis and selected based on the suspected source, clinical setting, knowledge or suspicion of causative organisms, and sensitivity patterns common to that specific inpatient unit. One regimen for septic shock of unknown cause includes gentamicin or tobramycin 5.1 mg/kg IV once/day plus a 3rd-generation cephalosporin (cefotaxime 2 g q 6 to 8 h or ceftriaxone 2 g once/day) or, if Pseudomonas is suspected, ceftazidime 2 g IV q 8 h. Vancomycin must be added if resistant staphylococci or enterococci are suspected. If there is an abdominal source, a drug effective against anaerobes (eg, metronidazole) should be included

### **Comments :** *when Temperature changed to 0.5 (0 was experimented previously)*

- **Temperature = 0 (Deterministic but Structurally Incomplete):** At temperature 0, the response was highly deterministic and closely grounded in the retrieved context, resulting in a comprehensive and protocol-oriented answer previously
- **Temperature = 0.5 (More Flexible but Less Stable):** At temperature 0.5, the model exhibits increased variability and attempts to provide more concrete drug-level details (specific antibiotics and dosages). *"if Pseudomonas is suspected, ceftazidime 2 g IV q 8 h. Vancomycin must be added if resistant staphylococci or enterococci are suspected. If there is an abdominal source, a drug effective against anaerobes (eg, metronidazole) should be included"*

# Question Answering using RAG-Finetuning



Parameters : `max_tokens=512,temperature=0.5,top_p=0.95,top_k=50`

## Question 2

What are the common symptoms for appendicitis, and can it be cured via medicine? If not, what surgical procedure should be followed to treat it??

### Response Received:

###Answer

The common symptoms for appendicitis include abdominal pain, anorexia, and abdominal tenderness. Appendicitis cannot be cured via medicine alone; instead, surgical removal of the appendix (appendectomy) is required for treatment.

**Comments :** when Temperature **changed to 0.5** (0 was experimented previously) and **number of token=512**

- **Temperature = 0 (Deterministic but Structurally Incomplete):** At temperature 0, the response was highly deterministic and closely grounded in the retrieved context, resulting in a comprehensive and protocol-oriented answer previously
- **Temperature = 0.5 (More Flexible but Less Stable):** At temperature 0.5, the model exhibits increased variability and attempts to provide very general answer
- **Not much impact** by increasing the number of **tokens to 512**

# Question Answering using RAG-Finetuning

Parameters : **max\_tokens=512,temperature=0.5,top\_p=0.95,top\_k=50**

## Question 3

What are the effective treatments or solutions for addressing sudden patchy hair loss, commonly seen as localized bald spots on the scalp, and what could be the possible causes behind it?

### Response Received :

Based on the context provided, the condition being described is Alopecia Areata. The effective treatments for this condition include topical, intralesional, or systemic corticosteroids, topical minoxidil, topical anthralin, topical immunotherapy (diphencyprone or squaric acid dibutylester), or psoralen plus ultraviolet A (PUVA). It is important to note that these treatments may have adverse effects and should be used under the guidance of a healthcare professional. The possible cause behind sudden patchy hair loss in Alopecia Areata is an autoimmune disorder affecting genetically susceptible people exposed to unclear environmental triggers.

**Comments :** when Temperature **changed to 0.5** (0 was experimented previously) and **number of token=512**

- **Temperature = 0 (Deterministic but Structurally Incomplete):** At temperature 0, the response was highly deterministic and closely grounded in the retrieved context, resulting in a comprehensive and protocol-oriented answer previously
- **Temperature = 0.5 (More Flexible but Less Stable):** At temperature 0.5, the model exhibits increased variability and attempts to provide multiple drugs and options to cure it “*The effective treatments for this condition include topical, intralesional, or systemic corticosteroids, topical minoxidil, topical anthralin, topical immunotherapy (diphencyprone or squaric acid ”*
- **Not much impact** by increasing the number of **tokens to 512**

# Question Answering using RAG-Finetuning

**Parameters:** max\_tokens=256,temperature=0.5,top\_p=0.95,top\_k=50

## Question 4.

What treatments are recommended for a person who has sustained a physical injury to brain tissue, resulting in temporary or permanent impairment of brain function?

### Response Received

From the context, it states that "Initial treatment consists of ensuring a reliable airway and maintaining adequate ventilation, oxygenation, and blood pressure. Surgery is often needed in patients with more severe injury to place monitors to track and treat intracranial pressure, decompress the brain if intracranial pressure is increased, or remove intracranial hematomas." Additionally, "Subsequently, many patients require rehabilitation." It also mentions that there is no specific treatment for TBI but supportive care should include preventing systemic complications due to immobilization, providing good nutrition, and preventing pressure ulcers. Therefore, the recommended treatments for a person who has sustained a physical injury to brain tissue, resulting in temporary or permanent impairment of brain function, are ensuring a reliable airway, maintaining adequate ventilation, oxygenation, and blood pressure, surgery if necessary, rehabilitation, preventing systemic complications due to immobilization, providing good nutrition, and preventing pressure ulcers.

**Comments :** when Temperature **changed to 0.5** (0 was experimented previously)

- **Temperature = 0 (Deterministic):** At temperature 0, the response was highly deterministic and closely grounded in the retrieved context, resulting in a comprehensive and protocol-oriented answer previously
- **Temperature = 0.5 (More Flexible but Less Stable):** At temperature 0.5, the model exhibits increased variability and attempts to provide multiple treatment options "*It also mentions that there is no specific treatment for TBI but supportive care should include preventing systemic complications due to immobilization, providing good nutrition, and preventing pressure ulcers.*"

# Question Answering using RAG-Finetuning



**Parameters:** max\_tokens=256, **temperature=0.5**, top\_p=0.95, top\_k=50

## Question 5

What are the necessary precautions and treatment steps for a person who has fractured their leg during a hiking trip, and what should be considered for their care and recovery?

### Response Received

The context provides information about the symptoms of a fracture, which include pain, swelling, ecchymosis, crepitation, deformity, and abnormal motion. It also mentions that treatment involves analgesics, immobilization, and sometimes surgery. The context specifically discusses the use of elastic bandages for initial injury care, elevation of the injured limb to minimize swelling, and periodic application of warmth after 48 hours to relieve pain and speed healing. It also mentions the importance of immobilizing joints proximal and distal to the injury, using a cast or splint depending on the length of immobilization required, and keeping the cast dry and inspecting the skin around it daily for signs of infection or irritation. The context also discusses potential complications of prolonged immobilization, such as stiffness, contractures, and muscle atrophy. It suggests that some rapidly healing injuries may benefit from early mobilization to minimize these complications. The context does not mention any specific precautions or treatment steps for a person who has fractured their leg during a hiking trip beyond the general information provided about fracture care and recovery. Therefore, based on the context

**Comments:** When Temperature **changed to 0.5** (0 was experimented previously)

- **Temperature = 0 (Deterministic but Truncated Output):** At temperature 0, the model has produces a deterministic, clinically relevant response in previous asks(only by RAG) that aligns well with the retrieved context and covers key precautions and treatment steps (elevation, immobilization, cast care).
- **Temperature = 0.5 (Verbose but Context-Diffuse):** At temperature 0.5, the model becomes more verbose and descriptive, largely summarizing the retrieved context rather than directly answering the question. While the information is accurate, the response lacks focus on the hiking-specific scenario and fails to synthesize clear precautions and recovery steps, weakening its practical value.

# Output Evaluation Prompt - Groundedness

You are tasked with rating AI generated answers to questions posed by users.

You will be presented a question, context used by the AI system to generate the answer and an AI generated answer to the question.

In the input, the question will begin with ###Question, the context will begin with ###Context while the AI generated answer will begin with ###Answer.

## **Evaluation criteria:**

The task is to judge the extent to which the metric is followed by the answer.

- 1 - The metric is not followed at all
- 2 - The metric is followed only to a limited extent
- 3 - The metric is followed to a good extent
- 4 - The metric is followed mostly
- 5 - The metric is followed completely

## **Metric:**

The answer should be derived only from the information presented in the context

## **Instructions:**

1. First write down the steps that are needed to evaluate the answer as per the metric.
2. Give a step-by-step explanation if the answer adheres to the metric considering the question and context as the input.
3. Next, evaluate the extent to which the metric is followed.
4. Use the previous information to rate the answer using the evaluation criteria and assign a score.

# Output Evaluation Prompt - Relevance

You are tasked with rating AI generated answers to questions posed by users.

You will be presented a question, context used by the AI system to generate the answer and an AI generated answer to the question.

In the input, the question will begin with ###Question, the context will begin with ###Context while the AI generated answer will begin with ###Answer.

## **Evaluation criteria:**

The task is to judge the extent to which the metric is followed by the answer.

- 1 - The metric is not followed at all
- 2 - The metric is followed only to a limited extent
- 3 - The metric is followed to a good extent
- 4 - The metric is followed mostly
- 5 - The metric is followed completely

## **Metric:**

**Relevance measures how well the answer addresses the main aspects of the question, based on the context. Consider whether all and only the important aspects are contained in the answer when evaluating relevance.**

## **Instructions:**

1. First write down the steps that are needed to evaluate the context as per the metric.
2. Give a step-by-step explanation if the context adheres to the metric considering the question as the input.
3. Next, evaluate the extent to which the metric is followed.
4. Use the previous information to rate the context using the evaluation criteria and assign a score.

# Output Evaluation- Groundedness Response

**Question 1:** What is the protocol for managing sepsis in a critical care unit?

**Parameters : Temp =0 and max\_tokens= 512** (Number of tokens are increased because important information was truncating)

## Steps to evaluate the answer:

1. Identify the key information related to managing sepsis in a critical care unit from the context.
2. Compare the identified information with the AI generated answer to check if the answer is derived only from the context.
3. Evaluate the extent to which the metric is followed.

## Explanation:

The context provides detailed information about managing critically ill patients in an ICU, including supportive care and patient monitoring. However, it specifically mentions sepsis management under the "Blood Tests" section. The AI generated answer correctly identifies the key components of managing sepsis in a critical care unit, which are administering antibiotics based on culture and sensitivity results, continuing antibiotic therapy for at least 5 days after shock resolves, draining abscesses, surgically excising necrotic tissues, and normalizing blood glucose. These components are directly derived from the context.

## Evaluation:

The metric is followed completely as the answer is derived only from the information presented in the context.

## Rating:

**Based on the evaluation criteria, I would rate the answer a 5 for following the metric completely.**

# Output Evaluation-Relevance Response

## **Question 1: What is the protocol for managing sepsis in a critical care unit?**

**Parameters : Temp =0 and max\_tokens= 512** (Number of tokens are increased because important information was truncating)

### **Steps to evaluate the context as per the relevance metric:**

1. Identify the main aspects of the question: managing sepsis in a critical care unit and the protocol for doing so.
2. Read through the context to understand the information provided about critical care units, patient monitoring, testing, and supportive care.
3. Look for specific information related to sepsis management in the context.
4. Determine if all important aspects of sepsis management are present in the context and if they are addressed in the AI generated answer.

### **The context adheres to the relevance metric considering the question as follows:**

1. The context provides information about critical care units, patient monitoring, testing, and supportive care, which is relevant to managing sepsis in a critical care unit.
2. The context includes specific information about sepsis management, such as the importance of prompt empiric therapy, antibiotic selection, and the role of drainage and surgical excision of septic foci.
3. All important aspects of sepsis management are present in the context, including administering antibiotics, considering culture and sensitivity results, draining abscesses, and normalizing blood glucose.

### **The metric is followed mostly as all important aspects of managing sepsis in a critical care unit are present in the context and addressed in the AI generated answer.**

Based on the evaluation criteria:

- 1 - The metric is not followed at all: 1
- 2 - The metric is followed only to a limited extent: 2
- 3 - The metric is followed to a good extent: 3
- 4 - The metric is followed mostly: 4
- 5 - The metric is followed completely: 5

Relevance response

**Rating for the context based on the relevance metric: 4.** (The context follows the relevance metric mostly as it includes all important aspects of managing sepsis in a critical care unit and addresses them in the AI generated answer.)

# Output Evaluation -Observations

**Question 1:** What is the protocol for managing sepsis in a critical care unit?

## Groundedness-Observations :

1. **Source Attestation Accuracy:** The evaluator successfully pinpointed the exact subsection ("Blood Tests") within the context, demonstrating a high ability to map specific claims to source locations.
2. **Strict Adherence to Negative Constraints:** The judge correctly confirmed that the AI did not hallucinate "outside knowledge" (like fluid resuscitation or vasopressors) that wasn't explicitly in your context.
3. **Logical Alignment:** The evaluator effectively used a "comparison check" step to ensure the 5-day post-shock antibiotic rule was derived strictly from the provided text.
4. **The metric is followed completely** as the answer is derived only from the information presented in the context **and 5/5 score is awarded**

## Relevance Observations

1. **Ambiguous Metric Application:** The evaluator confused the metric's target by rating the context for relevance instead of judging if the answer directly addressed the user's intent.
2. **Inconsistent Scoring Logic:** Despite stating that "all important aspects are present," **the judge only awarded a 4/5, failing to provide a specific reason for the point deduction.**
3. **Structural Reasoning Strength:** The judge excelled at deconstructing the user's prompt into key pillars (monitoring, testing, surgical excision) before searching for those matches in the response.

## Output Evaluation- Groundedness Response

**Question 2:** What are the common symptoms for appendicitis, and can it be cured via medicine? If not, what surgical procedure should be followed to treat it?

**Steps to evaluate the answer:**

1. Identify the information in the context related to appendicitis and its symptoms.
2. Determine if the AI generated answer includes only the information from the context.
3. Evaluate the extent to which the metric is followed.

**Explanation:**

The context provides information about appendicitis, including its common symptoms (abdominal pain, anorexia, and abdominal tenderness). The AI generated answer also mentions these same symptoms for appendicitis. Additionally, the answer states that appendicitis cannot be cured via medicine alone and that the standard treatment is surgical removal of the appendix through open or laparoscopic appendectomy. This information is directly from the context. Therefore, the AI generated answer adheres to the metric as it is derived only from the information presented in the context.

**Evaluation:**

The metric is followed completely.

**Rating:**

**Based on the evaluation criteria, I would rate the answer a 5 for following the metric completely.**

## Output Evaluation-Relevance Response

**Question 2:** What are the common symptoms for appendicitis, and can it be cured via medicine? If not, what surgical procedure should be followed to treat it?

### **Relevance response :**

#### **Steps to evaluate the context as per the relevance metric:**

1. Identify the main aspects of the question: common symptoms for appendicitis and whether it can be cured via medicine.
2. Determine if all important aspects are addressed in the context: The context mentions the common symptoms for appendicitis (abdominal pain, anorexia, and abdominal tenderness) and states that appendicitis cannot be cured via medicine alone. It also provides information about the standard treatment for appendicitis, which is surgical removal of the appendix through open or laparoscopic appendectomy.
3. Evaluate if only irrelevant aspects are missing: The context does not provide any irrelevant information that detracts from answering the question.

**The context adheres to the relevance metric as it addresses all important aspects of the question and does not include any irrelevant information.**

Rating the context using the evaluation criteria:

- 1 - The metric is not followed at all: 1
- 2 - The metric is followed only to a limited extent: 2
- 3 - The metric is followed to a good extent: 3
- 4 - The metric is followed mostly: 4
- 5 - The metric is followed completely: 5

Based on the evaluation criteria, **I would rate the context as a 5 because it follows the relevance metric completely.**

# Output Evaluation-Response -Observations

**Question 2** What are the common symptoms for appendicitis, and can it be cured via medicine? If not, what surgical procedure should be followed to treat it?

## Groundeness observations :

- Fact-Level Verification:** The evaluator successfully verified both clinical symptoms and treatment modalities (surgery vs. medicine) against the context, ensuring no external medical knowledge was smuggled in.
- Comprehensive Synthesis:** The judge correctly recognized that the answer's "negative claim" (cannot be cured by medicine) was grounded in the context's specific treatment descriptions.
- Process Transparency:** The evaluator followed a clear "identify and compare" workflow, which makes the final **score of 5 highly defensible and traceable.**

## Relevance observations :

- Target Misalignment:** The judge incorrectly evaluated the context's relevance to the question rather than the answer's relevance to the user.
- Negative Constraint Checking:** The evaluator added value by explicitly checking for "irrelevant information," ensuring the response stayed focused on the specific symptoms and cure constraints requested.
- Binary Aspect Matching:** The judge effectively broke the question into two distinct pillars (symptoms and medicinal cure) and confirmed both were satisfied before awarding a perfect score.
- Evaluation criteria followed**, and evaluator **rated the context as a 5 because it follows the relevance metric completely.**

## Output Evaluation- Groundedness Response

**Question 3:** What are the effective treatments or solutions for addressing sudden patchy hair loss, commonly seen as localized bald spots on the scalp, and what could be the possible causes behind it?

### Groundedness response :

#### **Steps to evaluate the answer:**

1. Identify the information in the context related to effective treatments or solutions for sudden patchy hair loss.
2. Check if the AI generated answer only includes the information from the context and does not add any new information that is not present in the context.
3. Verify if the possible causes mentioned in the AI generated answer are consistent with the context.

#### **Explanation:**

The AI generated answer adheres to the metric as it only includes the information from the context about the effective treatments for sudden patchy hair loss and mentions the possible cause, which is an autoimmune disorder, that is also mentioned in the context.

#### **Evaluation:**

The metric is followed completely.

#### **Rating:**

**Based on the evaluation criteria, I would rate the answer as 5 - The metric is followed completely.**

## Output Evaluation-Relevance Response

**Question 3:** What are the effective treatments or solutions for addressing sudden patchy hair loss, commonly seen as localized bald spots on the scalp, and what could be the possible causes behind it?

### **Relevance response :**

#### **Steps to evaluate the context as per the relevance metric:**

1. Identify the main aspects of the question: effective treatments or solutions for addressing sudden patchy hair loss and possible causes behind it.
2. Determine if the context provides information on the mentioned aspects.
3. Check if all important aspects are contained in the context and if any irrelevant information is present.

#### **The context adheres to the metric considering the question as the input because:**

1. The context mentions various treatments for sudden patchy hair loss, including topical, intralesional, or systemic corticosteroids, topical minoxidil, topical anthralin, topical immunotherapy (diphencyprone or squaric acid dibutylester), and psoralen plus ultraviolet A (PUVA).
2. The context also mentions possible causes behind sudden patchy hair loss, which could be an autoimmune disorder affecting genetically susceptible people exposed to unclear environmental triggers.
3. No irrelevant information is present in the context.

The metric is followed mostly because all important aspects of the question are addressed in the context and no irrelevant information is provided.

#### **Rating: 4 (The metric is followed mostly)**

## Output Evaluation Response –Observations

**Question 3:** What are the effective treatments or solutions for addressing sudden patchy hair loss, commonly seen as localized bald spots on the scalp, and what could be the possible causes behind it?

### Observations Groundedness :

- High Faithfulness to Scientific Terminology:** The evaluator correctly noted that the AI answer stuck to the specific medical treatments (corticosteroids, minoxidil, etc.) and the "autoimmune" cause mentioned in the context. This confirms the RAG system is successfully constrained to the provided document.
- Validation of Causal Links:** Observation 3 in the "Steps" section is particularly strong. The judge didn't just check for keywords; it checked if the relationship (Cause: autoimmune disorder hair loss) was consistent with the context. This prevents the model from incorrectly linking symptoms to the wrong causes.
- Potential "Self-Greeting" Bias:** Since the same model is evaluating itself, it tends to provide a very brief "Explanation." It confirms the presence of information but doesn't explicitly look for omissions. **For a score of 5,** the judge should ideally confirm that no extra medical advice (like "wear a hat" or "take vitamins") was added by the LLM from its own internal weights.

### Observations Relevance :

- Comprehensive Aspect Extraction:** The judge successfully identified that the user had a multi-part query: treatments AND causes. By breaking these into separate check-points, the judge ensures the retrieval engine didn't just return a list of drugs while ignoring why the hair loss is happening.
- Inconsistency in Scoring** (The "Strictness" Gap): Interestingly, **the judge gave a 4 for Relevance despite stating that "all important aspects are addressed" and "no irrelevant information is provided**
- Strictness on Context** Quality: Unlike the Groundedness evaluation, the Relevance evaluator looked at the Context rather than just the Answer. It listed specific treatments like **PUVA** and **anthralin**. This confirms that the retrieval step was successful in finding high-density information.

## Output Evaluation- Groundedness Response

**Question 4:** What treatments are recommended for a person who has sustained a physical injury to brain tissue, resulting in temporary or permanent impairment of brain function?

### **Groundedness response :**

#### **Steps to evaluate the answer:**

1. Identify the information in the context related to treatments for traumatic brain injury (TBI).
2. Determine if the AI generated answer includes only the information from the context regarding treatments for TBI.
3. Evaluate the extent to which the metric is followed.

#### **Explanation:**

The AI generated answer includes the following information derived from the context: ensuring a reliable airway and maintaining adequate ventilation, oxygenation, and blood pressure, surgery for patients with more severe injuries, maintaining adequate brain perfusion and oxygenation in the first few days after injury, rehabilitation, preventive measures for systemic complications due to immobilization, providing good nutrition, preventing pressure ulcers, and treatment for posttraumatic seizures.

The answer adheres to the metric as it is derived solely from the information presented in the context. Therefore, the metric is followed completely.

#### **Rating:**

**Based on the evaluation criteria, the AI generated answer receives a score of 5 since the metric is followed completely.**

# Output Evaluation-Relevance Response

**Question 4:** What treatments are recommended for a person who has sustained a physical injury to brain tissue, resulting in temporary or permanent impairment of brain function?

## **Relevance response :**

### **Steps to evaluate context as per relevance metric:**

1. Identify the main aspects of the question: treatments for a person with brain injury.
2. Determine if the context discusses these aspects in detail.
3. Check if all important aspects are covered and only relevant information is included.

### **The context adheres to the relevance metric considering the question as follows:**

1. The context mentions treatments for a person with brain injury, specifically discussing supportive care, initial treatment, surgery, and rehabilitation.
2. It provides detailed information about each aspect, including preventative measures, medications, and long-term management.
3. No irrelevant information is included in the context.

**The metric is followed mostly as the context covers all important aspects of the question** and only relevant information is provided.

**Therefore, I would rate the context a 4 on the evaluation criteria scale (followed mostly).**

# Output Evaluation-Response –Observations

**Question 4:** What treatments are recommended for a person who has sustained a physical injury to brain tissue, resulting in temporary or permanent impairment of brain function?

## Observations Groundedness :

- Granular Evidence Extraction:** The evaluator excelled at listing specific clinical interventions (airway management, perfusion, pressure ulcer prevention) to prove they were mapped directly from the source text.
- Temporal Logic Validation:** The judge correctly verified that the answer respected the context's timeline, distinguishing between "initial treatment" and "the first few days after injury."
- Strict Source Boundary: By awarding a 5,** the evaluator confirmed it resisted adding external TBI treatments (like specific ICP-lowering drugs or Glasgow Coma Scale references) not found in the context.

## Observations Relevance :

- Persistent Entity Confusion:** The evaluator continues to rate the context as the object of study ("Steps to evaluate context"), which ignores whether the AI's answer was actually helpful or concise for the user.
- Deduction Without Justification:** The judge assigned **a score of 4 ("Followed Mostly") despite explicitly stating that "all important aspects are covered" and "no irrelevant information is included," creating a logical contradiction.**
- Multimodal Treatment Mapping:** The judge effectively categorized the question's requirements into distinct phases—supportive care, surgery, and rehabilitation—to ensure a comprehensive relevance check.

# Output Evaluation-Groundedness Response

**Question 5:** What are the necessary precautions and treatment steps for a person who has fractured their leg during a hiking trip, and what should be considered for their care and recovery?

## **Groundedness response :**

### **Steps to evaluate the answer:**

1. Identify the information in the context related to precautions and treatment steps for a person with a fractured leg.
2. Compare each element of the AI generated answer to the corresponding information in the context.
3. Determine if the AI generated answer is derived solely from the context.

### **Explanation:**

The AI generated answer includes several elements related to precautions and treatment steps for a person with a fractured leg. These include keeping the cast dry, never putting an object inside the cast, inspecting the cast's edges and skin around it every day, applying lotion to any red or sore areas, padding any rough edges with soft material, seeking medical care if an odor emanates from within the cast or a fever develops, maintaining good hygiene, immobilization using a cast or splint, periodic application of warmth for pain relief after 48 hours, and considering early mobilization for rapidly healing injuries to minimize contractures and muscle atrophy.

Each of these elements can be found in the context. For example, keeping the cast dry is mentioned in the context as one of the instructions given to patients with casts. Similarly, periodic application of warmth for pain relief after 48 hours is mentioned in the context as a way to relieve pain and speed healing.

Therefore, based on this comparison, it can be concluded that the **AI generated answer is derived solely from the information presented in the context.**

### **Rating:**

**5 - The metric is followed completely.**

## Output Evaluation-Relevance Response

**Question 5:** What are the necessary precautions and treatment steps for a person who has fractured their leg during a hiking trip, and what should be considered for their care and recovery?

### **Relevance response :**

#### **Steps to evaluate context as per relevance metric:**

1. Identify the main aspects of the question: necessary precautions and treatment steps for a person with a fractured leg.
2. Read through the context to understand if it addresses these main aspects.
3. Check if all important information related to the question is present in the context.
4. Evaluate if any irrelevant information is included that may detract from the relevance of the answer.

#### **The context adheres to the metric considering the question as follows:**

1. The context provides detailed information about necessary precautions for a person with a fractured leg, such as keeping the cast dry, inspecting the skin around the cast, and seeking medical care if needed.
2. It also discusses treatment steps like immobilization using a cast or splint, application of ice, and sometimes surgery.
3. The context mentions complications that may arise from prolonged immobilization, such as stiffness, contractures, and muscle atrophy.
4. There is no irrelevant information in the context that detracts from the relevance of the answer.

The metric is followed mostly since all important aspects of the question are addressed in the context, and there is no irrelevant information included.

#### **Rating: 4 (The metric is followed mostly)**

# Output Evaluation-Response –Observations

**Question 5:** What are the necessary precautions and treatment steps for a person who has fractured their leg during a hiking trip, and what should be considered for their care and recovery?

## Observation Groundedness :

1. **Metric is followed completely:** Evaluator, did good comparison and concluded that **all fact are responded based on the context and rewarded score 5**
2. **High-Fidelity Comparison:** The evaluator successfully matched specific, granular care instructions (e.g., "padding rough edges" and "odor detection") to the source, ensuring no "common sense" hiking advice was hallucinated.
3. **Evidence-Based Explanation:** The evaluator provided concrete examples in the "Explanation" section (like the cast-dryness rule), which serves as a high-quality audit trail for the final score.

## Observation Relevance :

1. **Systemic Metric Failure:** The evaluator continues to evaluate the **context's potential** to answer the question rather than judging if the **actual AI answer** was helpful to the user who is stuck on a hiking trip.
2. **Unjustified Scoring Penalty:** The judge **awarded a 4/5 ("Followed Mostly")** despite explicitly stating that "**all important information is present**" and "**no irrelevant information is included,**" **showing a lack of critical differentiation between a 4 and a 5.**
3. **Functional Scoping:** The judge successfully identified **three distinct "care pillars"** (precautions, treatment, and recovery complications) to organize its assessment, even if the assessment **was directed at the wrong target.**



# APPENDIX

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

## Data Background and Contents

- **Dataset** : medical\_diagnosis\_manual.pdf
- **Total number of pages**: 4114
- **Chunk\_size**=512
- **Chunk\_Overlap**=20
- **Total number of Chunk**=8446

# Challenges Faced

- Older version of numpy was there in given notebook hence lot of dependency errors were popping and to fix it appropriate version **numpy==2.0.0** is installed
- Older version of the **lagchain== 0.3.27** , and it was not giving proper evaluation rating , hence appropriate version of the **lagchain ==1.0.0** was installed and because of this below line of code has to be changed as given code for executions was depreciated
- Retriever does not support the method **get\_relevant\_documents()** , hence new method **Invoke()** was called

```
rel_docs = retriever.get_relevant_documents("What are the common symptoms for appendicitis, and can it be cured via medicine? If not, what surgical procedure should be followed to treat it?") #Complete the code to pass the query→ Existing Code
```

```
rel_docs = retriever.invoke("What are the common symptoms for appendicitis, and can it be cured via medicine? If not, what surgical procedure should be followed to treat it?") #Complete the code to pass the query→ new code written
rel_docs
```

- **Invoke()** method takes only one parameter (**user\_input**) hence below code has to change

```
relevant_document_chunks = retriever.get_relevant_documents(query=user_input,k=3) → existing code
    relevant_document_chunks = retriever.with_config(
        {"search_kwargs": {"k": k}})
    .invoke(user_input) → New Code
```

- After making above code adjustment then I got perfect evaluation for question number including groundedness and Relevance rating for reference question number 1 evaluation output is available in next slide

# Output Evaluation-Response –Question 1-Reference slide

**Question: What is the protocol for managing sepsis in a critical care unit?**

Partial response due to the older version of the the **lagchain== 0.3.27**

Groundedness response

Steps to evaluate the answer:

1. Identify the information in the context related to managing sepsis in a critical care unit.
2. Compare the information identified in step 1 with the AI generated answer to determine if the answer is derived only from the context.
3. Evaluate the extent to which the metric (answer should be derived only from the information

Steps to evaluate the context as per the relevance metric:

1. Identify the main aspects of the question: managing sepsis in a critical care unit.
2. Read through the context to understand the information provided about critical care

Relevance response

Comment /Observation

**Great  
Learning**

**Power Ahead!**

