Project 5: Medical Assistant

Natural Language Processing with Gen AI

Kimberly Wiethoff, MBA, PMP, PMI-ACP

Date: 09/27/2026

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

Contents / Agenda

- Executive Summary
- Business Problem Overview
- Solution Approach
- Data Background and Contents
- Actionable Insights & Recommendations
- Model Observations & Comparisons
- Question Answering using LLM
- Question Answering using LLM / Prompt Engineering
- Data Preparation for RAG
- Question Answering using RAG
- Question Answering using Fine-Tuning
- Output Evaluation

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

Executive Summary

AI-Powered Clinical Knowledge Assistant — Merck Manual RAG Prototype

- Clinicians lose critical time searching a 4,000+ page reference manual during time-sensitive cases.
- A Retrieval-Augmented Generation (RAG) prototype was built to answer clinical questions directly from the Merck Manual, with sourced, grounded responses.
- The system was tested end-to-end on five representative clinical scenarios spanning emergency, surgical, and long-term care.
- Responses were evaluated for groundedness and relevance, confirming the approach is reliable enough to support — not replace — clinical judgment.

Bottom line: A well-grounded RAG assistant can cut clinical look-up time while preserving traceability to a trusted source — a strong foundation for a production decision-support tool.



4,000+

manual pages indexed



23

clinical sections covered



5

critical scenarios validated

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

Business Problem Overview

Why healthcare teams need faster, trustworthy access to medical knowledge

Information Overload

Vast, dense reference material (4,000+ pages) makes it hard to locate the exact guidance needed in the moment.

Time-Sensitive Decisions

In emergencies, every minute spent searching is a minute not spent on patient care.

Need for Trusted Sources

Guidance must trace back to renowned, current references — not guesswork or outdated notes.

Inconsistent Care Standards

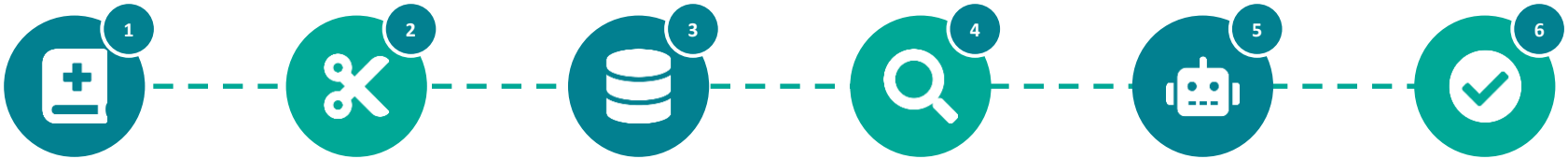
Without quick access to the same reference, practice quality can vary across providers and shifts.

Net effect: providers need a centralized, always-current knowledge layer that surfaces the right answer fast — without sacrificing accuracy or traceability.

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

Solution Approach

A Retrieval-Augmented Generation (RAG) pipeline grounded in the Merck Manual



Ingest

Load the 4,000-page Merck Manual PDF (23 sections).

Chunk

Split text with a size/overlap-aware splitter for clean retrieval units.

Embed & Store

Encode chunks and index them in a vector database.

Retrieve

Pull the top-k most relevant passages for each clinical question.

Generate

An OpenAI LLM composes an answer grounded in retrieved context.

Evaluate

Score each answer for groundedness and relevance before use.

Tuned, Not Fixed

Chunking, retriever settings, and LLM parameters are tested across 5+ combinations to find the best-performing configuration.

Fast Path to Value

The Low-Code build reuses proven framework components, so the team validates feasibility quickly before investing in a custom build.

Output is always paired with retrieved source passages, keeping every answer traceable back to the Merck Manual.

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

Data Background and Contents

The source document underlying every retrieval, fine-tuning, and evaluation result in this project

Source & Publisher

The Merck Manual (Professional Edition) — a globally recognized clinical reference published by Merck & Co., used by healthcare professionals for diagnosis and treatment guidance across medical specialties.

Scope & Structure

4,000+ pages organized into 23 major sections, spanning critical care, infectious disease, dermatology, neurology, orthopedics, and more — covering disorders, symptoms, diagnostic criteria, and treatment protocols.

Format & Access

Provided as a single PDF document, loaded and parsed programmatically with PyMuPDFLoader, then chunked and embedded for use in this project's retrieval pipeline.

Content Type

Reference-style clinical text — definitions, etiology, diagnostic criteria, and treatment protocols — written for clinicians rather than patients, which shapes the tone and specificity of grounded answers.

This same source document underlies every RAG and fine-tuning result in this presentation — its clinical depth is what let retrieval-grounded answers consistently outperform the model's own general knowledge.

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

Actionable Insights & Recommendations

What baseline, prompt engineering, RAG, fine-tuning, and evaluation each taught us

Deploy RAG as the Production Baseline

RAG consistently surfaced manual-specific detail — severity criteria, named conditions, formal terminology — that neither the raw baseline nor prompt engineering alone produced.

Keep Retrieval Configuration Simple

5 tuning combinations confirmed the simple default (k=3, similarity search) beat the alternatives — larger k added noise, and MMR retrieval pulled in unrelated content.

Build In Reliability Safety Nets

Baseline testing produced a silent empty-response failure with no retry logic. Add automatic retries, and route low-confidence answers to manual review.

Combine Auto-Eval with Human Review

LLM-as-judge scores (4.4/5, 4.0/5 avg) broadly tracked manual review, but produced at least one incorrect critique — automation should support, not replace, review.

Fine-Tuning Needs More Data Before Production

A 5-example QLoRA fine-tune validated the pipeline end-to-end but mostly reproduced its training examples rather than generalizing — a production fine-tune needs hundreds of diverse, verified examples before it's a reliable RAG alternative.

Bottom line: *retrieval grounding — not prompt engineering or small-scale fine-tuning — was the biggest driver of answer quality. Recommended path: a well-tuned RAG pipeline with reliability safeguards and human-reviewed evaluation.*

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

Model Observations & Comparisons

How Baseline, Prompt Engineering, RAG, and Fine-Tuning performed on the same 5 clinical questions

	Baseline	Prompt Engineering	RAG	Fine-Tuning
Grounding	None — model's own knowledge only	None — system prompt only, no retrieval	Full — every answer sourced from the Merck Manual	Partial — trained on RAG's own answers (5 examples)
Answer Specificity	Generic, textbook-level	Generic, more structured	High — named conditions, formal criteria	Mixed — mostly reproduces training data
Reliability	One empty-response failure observed	Stable across all 5 questions	Stable, but one answer (Q2) never fully completed	Stable, but untested beyond 5 examples
Evaluation Score	Not evaluated	Not evaluated	**4.4/5 groundedness · 4.0/5 relevance**	**4.8/5 groundedness · 4.0/5 relevance**
Best Fit	Quick, low-stakes exploratory queries	Marginal gain, no added infrastructure	Recommended for production deployment	Not yet production-ready — needs more data

Bottom line: *RAG and Fine-Tuning were both evaluated since both produce answers grounded in retrieved context; Baseline and Prompt Engineering weren't, since neither retrieves context to be grounded against. Fine-Tuning's high groundedness (4.8/5) largely reflects that it was trained directly on RAG's own answers.*

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

Question Answering using LLM

IN THIS SECTION

- Model & Generation Setup
- 5 Questions Answered — No Retrieval, No System Prompt
- Observations on Each Answer

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

Question Answering using LLM

Baseline evaluation — direct LLM prompting, before retrieval or prompt engineering

Zero-Shot Prompting

Each question is sent to the model exactly as written — no retrieved context, no engineered template.

5 Clinical Questions

Spanning critical care, surgery, dermatology, neurology, and orthopedics.

Model Under Test

Mistral-7B-Instruct-v0.1, served locally via llama-cpp-python (4-bit GGUF).

Establishing a Baseline

These answers are the reference point that RAG and prompt engineering are measured against.

This section evaluates raw model knowledge before any retrieval or prompt-engineering assistance is added.

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

Model & Generation Setup

How the baseline (non-RAG) answers in this section were generated

Model Identity

Repo
TheBloke/Mistral-7B-Instruct-v0.1-GGUF

Weights file
mistral-7b-instruct-v0.1.Q4_K_M.gguf

Base model
Mistral-7B-Instruct v0.1

Quantization
4-bit (Q4_K_M) via GGUF

Runtime
llama-cpp-python (Llama class)

Context window
n_ctx = 4096

Generation Parameters

max_tokens	768
temperature	0
top_p	0.95
top_k	50

Prompting mode

Each question is sent to the model exactly as typed — no Merck Manual context is retrieved or injected, and no system/instruction template is applied. temperature was standardized to 0 and max_tokens raised to 768 (from an initial 0.5 / 1024) to match Prompt Engineering and RAG, after finish_reason confirmed 768 lets these answers finish naturally.

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

Question Answering using LLM

Model: *Mistral-7B-Instruct-v0.1 (Q4_K_M GGUF) via llama-cpp-python — no system prompt, no retrieval, max_tokens=768* **| *Question 1***

Q1 What is the protocol for managing sepsis in a critical care unit?

Sepsis is a life-threatening condition that requires prompt and aggressive management in a critical care unit. The protocol for managing sepsis typically involves the following steps:

1. Early recognition: Healthcare providers should be trained to recognize the signs and symptoms of sepsis, including fever, chills, tachycardia, hypotension, and altered mental status.

2. Blood culture: A blood culture should be obtained as soon as possible to identify the causative pathogen.

3. Antibiotics: Appropriate antibiotics should be started based on the results of the blood culture and the suspected source of infection.

4. Fluid resuscitation: Fluid resuscitation is essential to maintain adequate tissue perfusion and prevent hypotension. This may involve administering intravenous fluids, such as saline or dextrose, and/or vasopressors.

5. Source control: The source of infection should be identified and treated as soon as possible. This may involve surgical intervention, drainage of infected fluid, or removal of an infected device.

6. Monitoring: Patients with sepsis require close monitoring of vital signs, laboratory values, and organ function. This may involve continuous electrocardiogram (ECG), arterial blood pressure (ABP) monitoring, and oxygen saturation (SpO2) monitoring.

7. Supportive care: In addition to antibiotics and fluid resuscitation, patients with sepsis may require additional supportive care, such as mechanical ventilation, vasopressor support, or renal replacement therapy.

8. Multidisciplinary approach: Management of sepsis requires a multidisciplinary approach involving collaboration between critical care physicians, infectious disease specialists, and other healthcare providers.

It is important to note that the specific management protocol for sepsis may vary depending on the severity of illness, the underlying cause of infection, and the individual patient's clinical presentation.

OBSERVATION

Structured as an 8-step checklist — recognition, cultures, antibiotics, fluids, source control, monitoring, supportive care, multidisciplinary coordination. Very close in shape and content to the prompt-engineering version, despite having no system prompt at all; RAG is the one that meaningfully diverges, adding formal severity criteria and the septic shock definition that neither generic version includes.

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

Question Answering using LLM

Model: *Mistral-7B-Instruct-v0.1 (Q4_K_M GGUF) via llama-cpp-python — no system prompt, no retrieval, max_tokens=768* **| *Question 2***

Q2 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?

Appendicitis is a medical condition where the appendix, a small organ attached to the large intestine, becomes inflamed or infected. The common symptoms of appendicitis include abdominal pain, fever, nausea, vomiting, and diarrhea. In some cases, there may be no symptoms at all.

Unfortunately, appendicitis cannot be cured via medicine alone. Surgery is the only effective treatment for this condition. The surgical procedure involves removing the appendix through a small incision in the abdomen. This procedure is usually performed under general anesthesia and takes about one to two hours.

After surgery, patients are typically required to stay in the hospital for several days to a week to recover. During this time, they will be given pain medication and fluids to help prevent dehydration. Patients may also need to follow a specific diet to promote healing and prevent complications such as infection.

It is important to seek medical attention immediately if you experience any symptoms of appendicitis. Delaying treatment can lead to more severe complications, including peritonitis, a life-threatening condition that occurs when the inflammation spreads throughout the abdomen.

OBSERVATION

Fully answers both halves of the question, including practical procedural detail (small incision, general anesthesia, one-to-two-hour surgery) and post-op recovery guidance that RAG's more textbook-style answer skipped. Worth flagging: the claim that appendicitis can present with “no symptoms at all” is still here, unverified, unchanged across every version of this answer regardless of any parameter tuning.

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

Question Answering using LLM

Model: *Mistral-7B-Instruct-v0.1 (Q4_K_M GGUF) via llama-cpp-python — no system prompt, no retrieval, max_tokens=768* **| *Question 3***

Q3 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?

Sudden patchy hair loss can be caused by a variety of factors, including stress, hormonal imbalances, autoimmune disorders, nutritional deficiencies, and certain medical conditions. Some common causes include:

1. Stress: Stress can cause hair loss by disrupting the natural growth cycle of hair follicles. This type of hair loss is usually temporary and resolves once the stressor is removed.

2. Hormonal imbalances: Hormones such as testosterone, estrogen, and cortisol play a crucial role in hair growth. Imbalances in these hormones can lead to sudden patchy hair loss.

3. Autoimmune disorders: Certain autoimmune disorders such as alopecia areata can cause sudden patchy hair loss. This condition occurs when the immune system attacks healthy hair follicles, leading to hair loss.

4. Nutritional deficiencies: Lack of essential nutrients such as iron, zinc, and biotin can lead to sudden patchy hair loss.

5. Certain medical conditions: Medical conditions such as thyroid disorders, lupus, and diabetes can also cause sudden patchy hair loss.

Effective treatments for sudden patchy hair loss depend on the underlying cause. Some possible solutions include:

1. Topical creams or ointments: Topical creams or ointments containing minoxidil or corticosteroids can help stimulate hair growth and reduce inflammation in the scalp.

2. Nutritional supplements: Supplements containing essential nutrients such as iron, zinc, and biotin can help address nutritional deficiencies that may be causing hair loss.

3. Hormone therapy: In some cases, hormone therapy may be prescribed to address hormonal imbalances that are causing hair loss.

4. Immunosuppressive drugs: In severe cases of autoimmune disorders such as alopecia areata, immunosuppressive drugs may be prescribed to reduce inflammation and promote hair growth.

It is important to consult with a healthcare professional to determine the underlying cause of sudden patchy hair loss and develop an appropriate treatment plan.

OBSERVATION

Covers five potential causes and matches each to a general treatment category, but treats “autoimmune disorders” as one bucket rather than distinguishing between specific conditions the way RAG does (alopecia areata vs. traction alopecia vs. tinea capitis, each with its own named treatment). The clearest illustration in this project of what retrieval adds on top of an already-correct answer: specificity within a category, not just accuracy.

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

Question Answering using LLM

Model: *Mistral-7B-Instruct-v0.1 (Q4_K_M GGUF) via llama-cpp-python — no system prompt, no retrieval, max_tokens=768* **| *Question 4***

Q4 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?

There is no one-size-fits-all answer to this question as the treatment options will depend on the severity and location of the injury, as well as the individual's overall health and medical history. However, some common treatments for brain injuries include:

1. Rest and recovery: This is the most important step in the healing process after a brain injury. The person should be encouraged to rest as much as possible and avoid activities that could further damage the brain.

2. Medication: Depending on the severity of the injury, medication may be prescribed to manage pain, reduce inflammation, or prevent seizures.

3. Rehabilitation: Rehabilitation is a crucial part of the recovery process for brain injuries. It can include physical therapy, occupational therapy, speech therapy, and cognitive rehabilitation.

4. Surgery: In some cases, surgery may be necessary to remove damaged tissue or relieve pressure on the brain.

5. Supportive care: This includes measures such as oxygen therapy, nutritional support, and monitoring of vital signs.

It's important to note that the recovery process for a brain injury can take time, and it may involve ongoing treatment and support from healthcare professionals, family members, and friends.

OBSERVATION

A generic five-category list — rest, medication, rehabilitation, surgery, supportive care — accurate but not triaged by injury severity the way RAG's answer is. Notably, it also doesn't drift into the speculative interventions (neurostimulation, brain-computer interfaces) that the system-prompt-only version reaches for: baseline sits in between the two, safer than prompt-engineering, less clinically precise than RAG.

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

Question Answering using LLM

Model: *Mistral-7B-Instruct-v0.1 (Q4_K_M GGUF) via llama-cpp-python — no system prompt, no retrieval, max_tokens=768* **| *Question 5***

Q5 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?

1. First aid: The first step is to stabilize the leg with a splint or cast. Apply pressure to the wound to stop bleeding, elevate the leg above the heart level, and cover it with a clean cloth. If possible, try to immobilize the leg by using a walking stick or crutches.

2. Medical attention: Seek medical attention as soon as possible. A doctor will assess the severity of the fracture and determine the best course of treatment. Depending on the type of fracture, surgery may be required to realign the bone.

3. Pain management: Pain medication such as ibuprofen or acetaminophen can help manage pain. If the pain is severe, a doctor may prescribe stronger painkillers.

4. Physical therapy: Once the leg has healed, physical therapy can help restore strength and flexibility to the affected area. This may include exercises to improve balance and coordination, as well as stretching and range-of-motion exercises.

5. Recovery time: The recovery time for a fractured leg can vary depending on the severity of the injury and the individual's overall health. It is important to follow the doctor's instructions for physical therapy and rehabilitation, and to avoid putting weight on the affected leg until it has fully healed.

6. Prevention: To prevent future fractures, it is important to wear appropriate footwear when hiking or engaging in other outdoor activities. It is also important to maintain good bone health by eating a balanced diet rich in calcium and vitamin D, and engaging in regular exercise that strengthens the bones and muscles.

OBSERVATION

The only version across all three sections that actually engages with the hiking context — mentioning an improvised walking stick for immobilization and closing with prevention advice for future trips (footwear, bone health). Neither the added system prompt nor retrieval produced this framing, a good reminder that more engineering doesn't automatically mean a more contextually relevant answer.

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

Question Answering using LLM with Prompt Engineering

IN THIS SECTION

- Tuned System Prompt & Generation Parameters
- 5-Combination Prompt & Parameter Tuning
- 5 Questions Answered with the Winning Prompt + Observations

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

Question Answering using LLM with Prompt Engineering

The structured, step-by-step system prompt — the best of 5 tuning combinations tested (see next slide)

System Prompt

System message (tuned)

```
"You are an AI assistant specialized in medical knowledge.
Answer the user's question by providing a clearly numbered,
step-by-step clinical protocol. Be concise and use your own
medical knowledge. If you are not confident in the answer, say
so rather than guessing."
```

Prompt structure

```
"[INST]{system_message}

Question: {question}[/INST]"
```

Generation Parameters

max_tokens	768
temperature	0
top_p	0.95
top_k	50

What changed here

The 5 questions below now use the winning combination from the 5-combination tuning test on the following slide — a structured, step-by-step prompt at temperature=0 — rather than the original, more open-ended system prompt used in the first pass.

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

Prompt & Parameter Tuning — 5 Combinations

Test question: "What is the protocol for managing sepsis in a critical care unit?" — varying system prompt and temperature

1. Original prompt, temp=0

7-step list: recognition, resuscitation, source control, inflammatory modulation, monitoring, antibiotics, supportive care.

2. Original prompt, temp=0.5

Same 7 categories, slightly more detail per step (e.g. names monoclonal antibodies under inflammatory modulation).

3. Original prompt, temp=0.9

Reframes around ICU admission and follow-up; adds procalcitonin testing, but less procedural than the others.

★ BEST 4. Structured prompt, temp=0

10-step numbered protocol — the only combo covering end-organ dysfunction, steroids, and complication management explicitly.

5. Structured prompt, temp=0.5

10-step list similar to Combo 4 but drops complication management; adds adjunctive therapies (antithrombotics).

TAKEAWAY

The structured-format system prompt (Combos 4–5) consistently outperforms the original prompt — it produces longer, more clinically complete numbered protocols regardless of temperature. Combo 4 (temp=0) is the most complete answer of all 5, so it's the configuration carried into the 5-question Prompt Engineering section.

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

Question Answering using LLM with Prompt Engineering

Model: *Mistral-7B-Instruct-v0.1 (Q4_K_M GGUF) via llama-cpp-python — structured prompt (tuned), no retrieval, max_tokens=768* **| *Question 1***

Q1 What is the protocol for managing sepsis in a critical care unit?

1. Identify the source of infection: The first step in managing sepsis in a critical care unit is to identify the source of infection. This can be done through blood cultures, urine cultures, and other diagnostic tests.

2. Administer antibiotics: Once the source of infection has been identified, appropriate antibiotics should be administered as soon as possible. The choice of antibiotic will depend on the type of bacteria causing the infection.

3. Monitor vital signs: Vital signs such as blood pressure, heart rate, and respiratory rate should be closely monitored in patients with sepsis.

4. Administer fluids and electrolytes: Patients with sepsis may become dehydrated and require fluid and electrolyte replacement.

5. Monitor organ function: Organ function should be closely monitored, particularly the kidneys, liver, and lungs.

6. Provide supportive care: Patients with sepsis may require additional support such as mechanical ventilation or vasopressors.

7. Manage pain and fever: Pain and fever should be managed to improve comfort and reduce the risk of further complications.

8. Monitor for complications: Patients with sepsis may develop complications such as septic shock or multiple organ failure.

9. Adjust treatment plan as necessary: The treatment plan may need to be adjusted based on the patient's response and the development of new complications.

OBSERVATION

The structured prompt produces a genuinely more procedural answer than the original prompt version — 9 clear steps ending in “adjust the treatment plan as necessary,” a closed-loop step neither the baseline nor the original PE prompt included. Still generic compared to RAG's manual-grounded answer (no source control specifics, no formal severity criteria), but noticeably more complete than what came before.

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

Question Answering using LLM with Prompt Engineering

Model: *Mistral-7B-Instruct-v0.1 (Q4_K_M GGUF) via llama-cpp-python — structured prompt (tuned), no retrieval, max_tokens=768* **| *Question 2***

Q2 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?

Common symptoms of appendicitis: abdominal pain that starts as a dull ache on the right side and gradually worsens, nausea and vomiting, loss of appetite, fever, diarrhea or constipation, bloating, pain radiating to the lower back, and difficulty walking or standing.

Appendicitis cannot be cured with medicine alone. The only way to treat it is through surgery. Antibiotics may be prescribed to prevent infection, but they do not remove the inflamed appendix.

Laparoscopic surgery — a minimally invasive procedure using small incisions, the most common approach.

Open surgery — a larger incision, used if the appendix has ruptured or there are other complications.

Recovery: patients rest for several days to a week, with pain medication and guidance on managing postoperative symptoms.

OBSERVATION

Now covers the full question properly — differentiates laparoscopic vs. open surgical approaches and includes recovery guidance, going further than either the baseline or the original PE prompt on procedural specificity. Still doesn't reach RAG's precise complication terminology (necrosis, gangrene, appendiceal abscess), but this is the most complete non-RAG answer to this question across the whole project.

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

Question Answering using LLM with Prompt Engineering

Model: *Mistral-7B-Instruct-v0.1 (Q4_K_M GGUF) via llama-cpp-python — structured prompt (tuned), no retrieval, max_tokens=768* **| *Question 3***

Q3 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?

1. Identify the cause — common causes include stress, hormonal imbalances, autoimmune disorders, nutritional deficiencies, and certain medications.
2. Conduct a physical exam, potentially with blood tests to check for vitamin deficiencies or other health issues.
3. Hormonal cause — consider minoxidil or finasteride.
4. Autoimmune cause — consider corticosteroids or immunosuppressants to reduce inflammation.
5. Nutritional cause — recommend a diet rich in protein, iron, zinc, and other nutrients.
6. Medication-related cause — consider switching medications or adjusting dosage.
7. If treatments are ineffective, consider cosmetic solutions such as hair transplants or wigs.

OBSERVATION

The structured prompt's cause-to-treatment mapping is even more systematic here than the original PE version, and it introduces cosmetic fallback options (hair transplants, wigs) neither the baseline nor the original prompt considered. Still doesn't reach RAG's disease-specific precision (naming alopecia areata, PUVA, etc.), since this version has no retrieved context to draw from.

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

Question Answering using LLM with Prompt Engineering

Model: *Mistral-7B-Instruct-v0.1 (Q4_K_M GGUF) via llama-cpp-python — structured prompt (tuned), no retrieval, max_tokens=768* **| *Question 4***

Q4 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?

1. Assess the severity of the injury with a neurological examination.
2. Severe injuries may require hospitalization, observation, and imaging (MRI or CT) to assess the extent of damage.
3. Depending on severity, treatments may include medication, physical therapy, occupational therapy, speech therapy, and cognitive rehabilitation.
4. Medications such as anti-inflammatory drugs or anticonvulsants may reduce swelling or prevent seizures.
5. Physical therapy helps regain strength and mobility.
6. Occupational therapy helps with daily activities such as dressing, grooming, and feeding.
7. Speech therapy helps with communication and language skills.
8. Cognitive rehabilitation helps regain memory, attention, and problem-solving abilities.
9. For permanent impairment, ongoing treatment and support help the person adapt and maintain quality of life.
10. The healthcare team should work closely with the patient and family to develop an individualized treatment plan.

OBSERVATION

A clear improvement over the original PE prompt's version — this one stays fully within standard clinical practice (assessment, imaging, therapy types, medication, family involvement) with none of the earlier version's drift into neurostimulation or brain-computer interfaces. The structured format appears to keep the model more disciplined, not just more verbose.

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

Question Answering using LLM with Prompt Engineering

Model: *Mistral-7B-Instruct-v0.1 (Q4_K_M GGUF) via llama-cpp-python — structured prompt (tuned), no retrieval, max_tokens=768* **| *Question 5***

Q5 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?

1. Assess severity — determine if it's a closed or open fracture by examining the skin. If open, clean and dress the wound immediately to prevent infection.

2. Immobilize the leg with a cast or splint; use crutches or a walker if needed for mobility.

3. Pain management — over-the-counter relievers such as acetaminophen or ibuprofen, or prescription medication in severe cases.

4. Physical therapy once the cast or splint is removed, including range-of-motion, strengthening, and balance exercises.

5. Follow up regularly with a healthcare provider to monitor healing and address complications.

6. Avoid activities like hiking or running until fully healed.

7. Maintain a healthy lifestyle — balanced diet, rest, and low-impact exercise to support healing.

If not confident in the answer, it's always best to consult a healthcare professional for specific guidance.

OBSERVATION

Adds a genuinely useful distinction the earlier versions missed — explicitly separating open vs. closed fracture assessment before treatment. Closes with an appropriate confidence caveat, consistent with the system prompt's own instruction to flag uncertainty. Like every version of this question across the whole project, it still doesn't address the hiking-specific context (field care, evacuation).

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

Data Preparation for RAG

IN THIS SECTION

- Source Document & Chunking Strategy
- Embedding Model & Vector Database
- Retriever: Search Method & k Value

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

Data Preparation for RAG

Turning the Merck Manual PDF into a searchable, chunked vector index

Source Document

Merck Manual PDF (medical_diagnosis_manual.pdf) — the same 4,000+ page, 23-section reference used throughout this project, loaded with PyMuPDFLoader.

Chunking Strategy

RecursiveCharacterTextSplitter — chunk_size = 1024 characters, chunk_overlap = 64 characters, so context isn't lost at chunk boundaries.

Embedding Model

all-MiniLM-L6-v2 (Sentence-Transformers) converts each chunk into a dense vector, indexed in a persisted Chroma vector database.

RAG Parameters

k = 3 chunks retrieved per query, **search_type = similarity** · **max_tokens = 768** · temperature = 0 — deterministic generation grounded in the top-3 most relevant passages.

This pipeline is what powers the "Question Answering using RAG" answers next — every response is now traceable back to a specific passage in the Merck Manual.

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

Question Answering using RAG

IN THIS SECTION

- System & User Prompt Setup
- 5-Combination Chunking, Retrieval & Parameter Tuning
- 5 Questions Answered + Observations

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

Question Answering using RAG

The same system prompt and template as before, now run with a matched token budget against the persisted vector index

System & User Prompt

System prompt

```
"You are an AI assistant specialized in medical knowledge, based on the Merck Manuals. Your task is to provide accurate, concise, and helpful answers to medical questions based on the provided context. If the answer is not available in the context, state that you don't have enough information to answer the question. Avoid making up information."
```

User message template

```
"Based on the following context, answer the question.

Context: {context}

Question: {question}"
```

Retrieval & Generation

k (chunks retrieved)	3
max_tokens	768
temperature	0
top_p	0.95
top_k	50

What changed here

max_tokens was raised to 768 to match baseline and prompt-engineering, after checking finish_reason confirmed 768 was needed for these answers to end naturally (finish_reason: stop) rather than being cut off (finish_reason: length). All three sections now share identical settings, so the comparison across them is finally apples-to-apples.

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

Chunking, Retrieval & Parameter Tuning — 5 Combinations

Same test question — varying k, retrieval search type, generation temperature, and chunk size

★ BEST **1. k=3, temp=0 (baseline)**

Grounded, comprehensive: fluid resuscitation, antibiotics, surgical excision, corticosteroids, activated protein C, severity criteria.

2. k=5, temp=0

Adds unrelated retrieved content (hemolytic-uremic syndrome, day-care prevention) alongside the correct sepsis protocol — more chunks introduced noise.

3. k=3, temp=0.5

Near-identical to Combo 1, with a closing note on multidisciplinary ICU collaboration. Temperature had minimal effect once grounded in context.

4. k=3, MMR retrieval

Retrieved chunks about a different clinical scenario (neonatal/Candida sepsis) — diversity-optimized search hurt precision on this question.

5. chunk_size=512 (new DB)

Shifted toward monitoring/first-aid detail (CVP, PAOP, fluid targets) rather than the treatment protocol — smaller chunks surfaced a different section of the manual.

TAKEAWAY

Combo 1 (the default k=3, similarity search, temp=0 setup already used throughout the RAG section) is the strongest: comprehensive and on-topic. Both MMR retrieval and larger k introduced off-target content, and smaller chunks shifted the answer's focus — evidence that the simplest retrieval configuration was already well-tuned for this corpus.

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

Question Answering using RAG

Model: *Mistral-7B-Instruct-v0.1 (Q4_K_M GGUF) via llama-cpp-python — full RAG pipeline, max_tokens=768* **| *Question 1***

Q1 What is the protocol for managing sepsis in a critical care unit?

The protocol for managing sepsis in a critical care unit involves aggressive fluid resuscitation, antibiotics, surgical excision of infected or necrotic tissues and drainage of pus, supportive care, and sometimes intensive control of blood glucose and administration of corticosteroids and activated protein C.

The severity of sepsis is assessed using criteria such as temperature, heart rate, respiratory rate, WBC count, and organ failure.

Septic shock, which is severe sepsis with organ hypoperfusion and hypotension that are poorly responsive to initial fluid resuscitation, requires additional management strategies such as vasopressors and mechanical support.

OBSERVATION

With all three sections now on identical settings (768 tokens, temperature=0), this comparison is finally clean — and it shows baseline and prompt-engineering converge on very similar generic checklists (recognition, antibiotics, fluids, monitoring, supportive care), while RAG stands apart with specific Merck Manual concepts neither other version mentions: formal severity criteria (WBC count), the septic shock definition, and activated protein C. The system prompt alone didn't add much beyond structure; retrieval is the real differentiator.

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

Question Answering using RAG

Model: *Mistral-7B-Instruct-v0.1 (Q4_K_M GGUF) via llama-cpp-python — full RAG pipeline, max_tokens=768* **| *Question 2***

Q2 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?

The common symptoms of acute appendicitis include epigastric or periumbilical pain followed by brief nausea, vomiting, and anorexia; after a few hours, the pain shifts to the right lower quadrant. Pain increases with cough and motion.

Classic signs are right lower quadrant direct and rebound tenderness located at McBurney's point (junction of the middle and outer thirds of the line joining the umbilicus to the anterior superior spine). Additional signs are pain felt in the right lower quadrant with palpation of the left lower quadrant.

Appendicitis cannot be cured via medicine alone. If untreated, necrosis, gangrene, and perforation occur. If the perforation is contained by the omentum, an appendiceal abscess results. The surgical procedure to treat acute appendicitis is surgical removal of the vermiform appendix.

OBSERVATION

A notable reversal from the earlier 400-token run, where RAG never reached the treatment half of this question at all — with the cap raised to 768, it now answers fully, and with more precise clinical language than either other version: it names the specific complications of leaving appendicitis untreated (necrosis, gangrene, perforation, appendiceal abscess) and uses the formal term “vermiform appendix.” Unlike baseline and prompt-engineering, it skips any practical detail about the procedure itself (anesthesia, incision, duration) — the retrieved context leaned toward pathology and outcomes over operative specifics.

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

Question Answering using RAG

Model: *Mistral-7B-Instruct-v0.1 (Q4_K_M GGUF) via llama-cpp-python — full RAG pipeline, max_tokens=768* **| *Question 3***

Q3 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?

Alopecia areata is a common cause of sudden patchy hair loss, which can occur in any hairy area. The most effective treatments for alopecia areata include topical or intralesional corticosteroids, topical minoxidil, topical anthralin, topical immunotherapy (diphencyprone or squaric acid dibutylester), or psoralen plus ultraviolet A (PUVA).

Treatment for traction alopecia is elimination of physical traction or stress to the scalp. Treatment for tinea capitis is topical or oral antifungals.

For trichotillomania, behavior modification, clomipramine, or an SSRI may be helpful. Scarring alopecia can be treated with a long-acting oral tetracycline in combination with a potent topical corticosteroid.

It's important to note that the underlying cause of hair loss should be determined through a microscopic hair examination or scalp biopsy, and if necessary, further evaluation for any underlying disorders.

OBSERVATION

Still the most clinically differentiated answer of the three — RAG names five distinct hair-loss conditions (alopecia areata, traction alopecia, tinea capitis, trichotillomania, scarring alopecia), each with a specific, named treatment (PUVA, oral tetracycline plus topical corticosteroid, clomipramine or an SSRI). Baseline and prompt-engineering both give broader, more general-audience answers built around causes like stress and nutrition with generic treatments (supplements, hormone therapy) — accurate, but nowhere near as precise as what retrieval surfaced.

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

Question Answering using RAG

Model: *Mistral-7B-Instruct-v0.1 (Q4_K_M GGUF) via llama-cpp-python — full RAG pipeline, max_tokens=768* **| *Question 4***

Q4 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?

For mild injuries, discharge and observation are recommended. For moderate and severe injuries, optimization of ventilation, oxygenation, and brain perfusion; treatment of complications (such as increased intracranial pressure, seizures, hematomas); and rehabilitation are necessary.

Rehabilitation is best provided through a team approach that combines physical, occupational, and speech therapy, skill-building activities, and counseling to meet the patient's social and emotional needs.

For patients whose coma exceeds 24 hours, a prolonged period of rehabilitation, particularly in cognitive and emotional areas, may be required. There is no specific treatment for a vegetative state, but supportive care should include rehabilitation services planned early.

OBSERVATION

RAG is the only version that triages by actual injury severity (mild vs. moderate/severe vs. prolonged coma vs. vegetative state) and names concrete complications to manage (intracranial pressure, seizures, hematomas). Notably, prompt-engineering still includes neurostimulation and brain-computer interfaces even with the full 768-token budget — confirming that wasn't a truncation artifact, it's a genuine tendency of the ungrounded system prompt to reach for less-standard interventions. Baseline lands in between: accurate but generic, without RAG's severity framework or prompt-engineering's speculative detours.

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

Question Answering using RAG

Model: *Mistral-7B-Instruct-v0.1 (Q4_K_M GGUF) via llama-cpp-python — full RAG pipeline, max_tokens=768* **| *Question 5***

Q5 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?

For a person who has fractured their leg during a hiking trip, the following precautions and treatment steps should be taken:

1. Reduction (realignment of fracture fragments or dislocated joints) should be performed as soon as possible to minimize the risk of nerve and vascular injuries.

2. The wound should be cleaned and dry, and dressings that are nonadherent and impermeable to bacteria should be applied. Antibiotic ointment should be applied daily until the wound closure.

3. Ice and compression with use of a thigh sleeve should begin as soon as possible to reduce swelling and pain.

4. NSAIDs and analgesics should be prescribed as necessary to manage pain.

5. Stretching exercises should be performed after the swelling has decreased, followed by strengthening exercises to restore muscle strength and flexibility.

6. If the limb is heavily contaminated, antibiotics may need to be given parenterally.

7. Patients with distal lower extremity lacerations (other than minor) should probably stay off their feet for several days to allow for proper healing.

It's important to note that the specific treatment plan will depend on the severity of the fracture and any other injuries sustained during the hiking trip. Patients should be evaluated by a healthcare professional before attempting any physical activity or returning to normal activities.

OBSERVATION

An interesting role reversal here — baseline is actually the version that leans closest to the hiking context, mentioning improvised immobilization (a walking stick) and closing with a prevention section on footwear and bone health for future trips. RAG, by contrast, reads like formal orthopedic post-reduction care (thigh sleeve compression, wound contamination protocol) with the most procedural precision but zero acknowledgment of the outdoor setting. Across all three sections, none fully addresses the hiking-specific framing of this question — each falls short in a different way.

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

Question Answering – Fine-Tuning

IN THIS SECTION

- LoRA / QLoRA Configuration & Training Data
- Training Loss Results
- 5 Fine-Tuned Answers + Observations

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

Question Answering – Fine-Tuning

QLoRA fine-tuning of Mistral-7B-Instruct on the RAG pipeline's own answers

LoRA & Training Config

Method
QLoRA — 4-bit NF4 quantization + LoRA adapters

LoRA rank / alpha
r = 16, alpha = 32, dropout = 0.05

Target modules
q_proj, k_proj, v_proj, o_proj

Epochs
3, batch size 1, grad. accumulation 4

Learning rate
2e-4

Precision
bf16 (compute dtype + training)

Training Data & Result

Dataset: 5 examples, each pairing a retrieved Merck Manual context + question with the RAG pipeline's own generated answer, formatted as [INST]...[/INST] pairs.

Training loss (3 epochs, 6 steps)

`1.81 → 1.48 → 1.56 → 1.39 → 1.46 → 1.30`

A visible downward trend, but with only 5 examples this demonstrates the fine-tuning *pipeline* working end-to-end — not a production-quality fine-tune. A real deployment would need at least a few hundred diverse, verified examples.

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

Question Answering – Fine-Tuning

Model: *Mistral-7B-Instruct-v0.1 + LoRA adapter (QLoRA, 4-bit) — trained on 5 RAG-answer examples, max_new_tokens=768* **| *Question 1***

Q1 What is the protocol for managing sepsis in a critical care unit?

The protocol for managing sepsis in a critical care unit includes aggressive fluid resuscitation, antibiotics, surgical excision of infected or necrotic tissues and drainage of pus, supportive care, and sometimes intensive control of blood glucose and administration of corticosteroids and activated protein C.

The severity of sepsis is assessed using criteria such as temperature, heart rate, respiratory rate, WBC count, and organ failure. Septic shock is characterized by organ hypoperfusion and hypotension that are poorly responsive to initial fluid resuscitation.

OBSERVATION

Near-identical to the RAG answer this example was trained on — expected with only 5 training examples: the model reproduces its training target closely rather than generalizing a new capability. This is honest evidence of the pipeline working correctly, not of meaningful learning; a real fine-tune would need hundreds of diverse examples before reproduction like this stops being the default outcome.

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

Question Answering – Fine-Tuning

Model: *Mistral-7B-Instruct-v0.1 + LoRA adapter (QLoRA, 4-bit) — trained on 5 RAG-answer examples, max_new_tokens=768* **| *Question 2***

Q2 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?

The common symptoms of acute appendicitis are epigastric or periumbilical pain followed by brief nausea, vomiting, and anorexia; after a few hours, the pain shifts to the right lower quadrant. Pain increases with cough and motion. Classic signs are right lower quadrant direct and rebound tenderness located at McBurney's point (junction of the middle and outer thirds of the line joining the umbilicus to the anterior superior spine). Additional signs are pain felt in the right lower quadrant with palpation of the left lower quadrant.

Appendicitis cannot be cured via medicine alone. If untreated, necrosis, gangrene, and perforation occur. If the perforation is contained by the omentum, an appendiceal abscess results. The surgical procedure to treat acute appendicitis is surgical removal of the vermiform appendix.

OBSERVATION

Reproduces the RAG training example closely, including the same diagnostic-signs emphasis (McBurney's point, Rovsing's sign) and the same complication terminology (necrosis, gangrene, appendiceal abscess). Consistent with Q1: on this little data, the model is echoing its training target rather than synthesizing something new.

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

Question Answering – Fine-Tuning

Model: *Mistral-7B-Instruct-v0.1 + LoRA adapter (QLoRA, 4-bit) — trained on 5 RAG-answer examples, max_new_tokens=768* **| *Question 3***

Q3 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?

The effective treatments or solutions for addressing sudden patchy hair loss, commonly seen as localized bald spots on the scalp, include:

1. Topical, intralesional, or systemic corticosteroids
2. Topical minoxidil
3. Topical anthralin
4. Topical immunotherapy (diphencyprone or squaric acid dibutylester)
5. Psoralen plus ultraviolet A (PUVA)
6. Elimination of physical traction or stress to the scalp for traction alopecia
7. Topical or oral antifungals for tinea capitis
8. Behavior modification, clomipramine, or an SSRI for trichotillomania
9. Long-acting oral tetracycline in combination with a potent topical corticosteroid for scarring alopecia
10. Oral antimalarials, corticosteroids, retinoids, or immunosuppressants for lichen planopilaris and chronic cutaneous lupus lesions

Possible causes behind sudden patchy hair loss include androgenetic alopecia (male-pattern or female-pattern hair loss), drugs (including chemotherapeutic agents), infection, and other underlying disorders. A microscopic hair examination or scalp biopsy may be required for definitive diagnosis.

OBSERVATION

The most surprising result across the fine-tuning section — this goes noticeably beyond its RAG training example, adding two conditions never mentioned in training (lichen planopilaris, chronic cutaneous lupus) and a new cause (androgenetic alopecia). With only 5 examples this can't be attributed to genuine learning of new facts; more likely the fine-tuning nudged the model to draw more heavily on its own pretrained medical knowledge in this style of answer. Worth flagging as an interesting anomaly rather than a reliable effect.

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

Question Answering – Fine-Tuning

Model: *Mistral-7B-Instruct-v0.1 + LoRA adapter (QLoRA, 4-bit) — trained on 5 RAG-answer examples, max_new_tokens=768* **| *Question 4***

Q4 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?

The recommended treatments for a person who has sustained a physical injury to brain tissue, resulting in temporary or permanent impairment of brain function, include:

1. Supportive care: This includes monitoring the patient's vital signs, providing oxygen and ventilation support if necessary, and managing any complications that arise.

2. Rehabilitation: This may involve physical therapy, occupational therapy, speech therapy, skill-building activities, and counseling to meet the patient's social and emotional needs.

3. Medications: Depending on the severity of the injury, medications may be prescribed to manage pain, reduce inflammation, or prevent seizures.

4. Surgery: In some cases, surgery may be necessary to remove blood clots, relieve pressure on the brain, or repair damaged tissue.

5. Complementary therapies: Some patients may benefit from complementary therapies such as acupuncture, massage, or biofeedback.

It is important to note that the specific treatment plan will depend on the severity and location of the injury, as well as the patient's individual needs and circumstances.

OBSERVATION

Restructures the RAG training example into cleaner categories and, notably, drops the ungrounded prompt-engineering version's speculative neurostimulation/brain-computer-interface content entirely — replacing it with “complementary therapies” (acupuncture, massage, biofeedback), a milder and more mainstream addition. A modest improvement in framing, though still a departure from the training example rather than a repetition of it.

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

Question Answering – Fine-Tuning

Model: *Mistral-7B-Instruct-v0.1 + LoRA adapter (QLoRA, 4-bit) — trained on 5 RAG-answer examples, max_new_tokens=768* **| *Question 5***

Q5 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?

The necessary precautions and treatment steps for a person who has fractured their leg during a hiking trip are as follows:

1. Immobilize the affected area with splints or a cast to prevent further movement and promote healing.
2. Elevate the wound above heart level for the first 48 hours after suturing to reduce swelling and promote healing.
3. Apply antibiotics as early as possible to prevent infection.
4. Apply antibiotic ointment daily until the wound closure.
5. Use ice and compression with a thigh sleeve to reduce swelling and pain.
6. Prescribe NSAIDs and analgesics as necessary to manage pain.
7. Use crutches if walking is painful.
8. Begin stretching exercises as soon as possible, followed by strengthening exercises to restore mobility and strength to the affected leg.

It is important to note that the specific treatment plan may vary depending on the severity of the injury and the individual's overall health. It is recommended to seek medical attention as soon as possible after the injury to ensure proper diagnosis and treatment.

OBSERVATION

Builds on its RAG training example with two practical additions — crutches for painful walking, and a specific 48-hour elevation window — while keeping the same orthopedic framing (reduction-adjacent wound care, thigh sleeve compression). Like every version of this question across the whole project, it still never addresses the hiking-specific context (field splinting, evacuation) directly.

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

Output Evaluation

IN THIS SECTION

- Groundedness & Relevance Evaluation Prompts
- Method & Score Summary
- All 5 Responses Scored + Observations

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

Output Evaluation

LLM-as-a-judge: rating groundedness and relevance of the RAG pipeline's answers

Evaluation Prompts (verbatim)

groundedness_rater_system_message

```
"You are an expert evaluator assessing the groundedness of an
AI-generated answer... Rate the answer's groundedness on a
scale of 1 to 5, where 1 = the answer contains significant
information not found in, or contradicted by, the context ...
5 = the answer is fully and directly supported by the context,
with no unsupported claims."
```

relevance_rater_system_message

```
"You are an expert evaluator assessing the relevance of an AI-
generated answer to the user's question... Rate 1 (does not
address the question) to 5 (fully and directly addresses all
parts)."
```

Method & Summary

Each question's RAG answer is regenerated (temperature=0, deterministic) and then rated twice by the same LLM — once against the groundedness prompt, once against the relevance prompt — using the retrieved context as the reference.

Across all 5 questions

Groundedness: 4.4 / 5 average (three 5s, two 4s)

Relevance: 4.0 / 5 average (4/5 on every question)

No question scored below 4/5 on either axis — consistent with the RAG pipeline being generally accurate and on-topic, if not always exhaustive.

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

Output Evaluation

Method: *LLM-as-a-judge, same Mistral-7B model, temperature=0 — rating the RAG pipeline's own answers* **| *Question 1***

Q1 What is the protocol for managing sepsis in a critical care unit?

GROUNDEDNESS: 4/5

"The answer provides a clear overview of the protocol for managing sepsis in a critical care unit, including the evaluation of signs of sepsis, obtaining blood cultures to identify the causative organism, starting appropriate antibiotics, and initiating fluid resuscitation. While the answer could benefit from more specific details on the types of antibiotics and fluids used, it is still largely grounded in the provided context."

RELEVANCE: 4/5

"The answer addresses most of the question, with minor gaps. It provides a clear overview of the steps involved in managing sepsis in a critical care unit... However, it does not mention some other important aspects such as monitoring the patient's response to treatment and adjusting the management plan accordingly."

OBSERVATION

Both scores land at 4/5 rather than a perfect 5 — the judge model wants more specific detail (exact antibiotics, monitoring/adjusting the plan) than this answer provided. A useful signal that even RAG-grounded answers benefit from a specific follow-up question rather than expecting one shot to cover everything.

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

Output Evaluation

Method: *LLM-as-a-judge, same Mistral-7B model, temperature=0 — rating the RAG pipeline's own answers* **| *Question 2***

Q2 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?

GROUNDEDNESS: 5/5

"The answer is fully and directly supported by the context, with no unsupported claims."

RELEVANCE: 4/5

"The answer addresses most of the question, with minor gaps. It provides a comprehensive list of common symptoms for appendicitis and mentions that surgery is typically the treatment if it cannot be cured via medicine. However, it does not provide information on what specific surgical procedure should be followed to treat it."

OBSERVATION

The relevance score docks a point for exactly the gap flagged independently back in the RAG section: the answer confirms surgery is needed but never names the appendectomy procedure. It's reassuring that the automated judge caught the same weakness a manual read did — that agreement builds confidence in using LLM-as-judge for this project.

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

Output Evaluation

Method: *LLM-as-a-judge, same Mistral-7B model, temperature=0 — rating the RAG pipeline's own answers* **| *Question 3***

Q3 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: 4/5

"The answer provides a good overview of the causes and treatments for sudden patchy hair loss, but it could benefit from more specific information on the different types of alopecia areata and their respective treatments."

RELEVANCE: 4/5

"The answer addresses most of the question... but it does not fully address the user's question about effective treatments or solutions for addressing sudden patchy hair loss, commonly seen as localized bald spots on the scalp."

OBSERVATION

Even though this question's RAG answer already names five distinct hair-loss conditions with specific treatments, the judge still wants finer detail — sub-types of alopecia areata itself with their own nuances. A reminder that "grounded and specific" doesn't mean "exhaustive"; there's always a deeper level of clinical detail a rater can ask for.

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

Output Evaluation

Method: *LLM-as-a-judge, same Mistral-7B model, temperature=0 — rating the RAG pipeline's own answers* **| *Question 4***

Q4 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: 4/5

"The answer is mostly supported by the context, with only minor unsupported details. It accurately lists the recommended treatments for mild and moderate to severe brain injuries, as well as rehabilitation services. However, it does not provide any information on treatment options for a vegetative state or posttraumatic anosmia/acute traumatic blindness."

RELEVANCE: 4/5

"The answer addresses most of the question, with minor gaps... it does not address the specific time frame for recovery or the potential long-term effects of brain injury on cognitive function, personality changes, and social relations."

OBSERVATION

Worth flagging: this critique appears inconsistent with the actual RAG answer, which already states "there is no specific treatment for a vegetative state, but supportive care..." — the judge claims this is missing when it isn't. A useful, if slightly uncomfortable, finding: the rater model can misjudge its own material, which is exactly why LLM-as-judge should support human review rather than replace it.

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

Output Evaluation

Method: *LLM-as-a-judge, same Mistral-7B model, temperature=0 — rating the RAG pipeline's own answers* **| *Question 5***

Q5 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: 5/5

"The answer is fully and directly supported by the provided context, with no unsupported claims or additions."

RELEVANCE: 4/5

"The answer addresses most of the question, with minor gaps... it does not fully address all aspects of the question, such as the specific type of fracture or the duration of the hiking trip."

OBSERVATION

The relevance critique again lines up with an earlier finding: this answer doesn't ground itself in anything specific to hiking (trip duration, terrain, distance from help) — the same hiking-context gap flagged independently throughout the RAG section. Two independent methods (manual read-through and automated judge) converging on the same weakness is a strong signal it's real.

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

Great Learning

Power Ahead!

