

Why RAG Is the Superior Approach for a Medical AI Assistant

Medical Assistant - GenAI for Business Applications

December 8, 2025

By: Jordan Plank

Executive Summary

Key Insights:

- RAG provides clinically grounded answers by pulling directly from verified medical manuals, significantly reducing hallucinations.
- LLMs, even with prompt engineering, cannot guarantee factual accuracy and often produce unsafe or unverified medical content.
- RAG offers faster updates (no retraining required) and supports traceable, protocol-aligned outputs preferred by clinicians.
- Retrieval quality (k value, chunking, indexing) is the main driver of RAG performance.

Recommendations:

- Adopt a RAG-first architecture for all medical AI use cases to ensure accuracy and safety.
- Implement a clinical knowledge governance process to maintain up-to-date, version-controlled medical content.
- Tune retrieval parameters and enforce explicit fallback behavior when information is not found.
- Use targeted prompt engineering only for structure and clarity — not factual accuracy.

Executive Takeaway: RAG is the most reliable and clinically safe approach for a medical AI assistant, outperforming standard LLM methods in accuracy, safety, and operational efficiency.

Contents / Agenda

- Business Problem and Approach
- Question Answering using LLM
- Question Answering using LLM with Prompt Engineering
- Data Preparation for RAG
- Question Answering using RAG
- Output Evaluation

Business Problem Overview

Healthcare professionals are overwhelmed by rapidly expanding medical knowledge, complex diagnostic requirements, and the need to make fast, accurate decisions—often in high-pressure environments such as emergency care and critical care units. Clinicians struggle to manually search through medical manuals, research papers, and fragmented information sources, resulting in:

- **Information overload** that slows down clinical decision-making
- **Difficulty accessing trusted, up-to-date medical references** when time is limited
- **Risk of inconsistent diagnoses and treatment plans** across providers
- **Reduced efficiency and potential delays in patient care**, especially in emergencies

Examples of frequently asked clinical queries—ranging from sepsis protocols to surgical decision pathways to dermatological or neurological assessments—highlight the dependency on authoritative, domain-specific information. Traditional methods of retrieving this knowledge are too slow, too manual, and not optimized for point-of-care use.

Solution Methodology

1. Knowledge Ingestion

- Convert the 4,000-page Merck Manual PDF into clean, structured text.
- Split content into meaningful chunks and generate embeddings for semantic retrieval.

2. RAG Pipeline Construction

- Build a vector search system to retrieve the most relevant manual sections for any clinical question.
- Use an LLM to generate concise, grounded responses directly tied to the retrieved content, with a safe fallback when information is unavailable.

3. Clinical Use-Case Testing

- Validate the system with representative questions from key specialties (critical care, surgery, dermatology, neurology).
- Assess accuracy, clarity, and medical appropriateness of the responses.

4. Evaluation & Optimization

- Measure groundedness, precision/relevance, and completeness.
- Tune retrieval settings (e.g., top-k) and refine chunks to maximize answer quality.

5. Prototype Delivery

- Build a functional RAG prototype that accepts clinician queries, retrieves the best evidence, and returns clear, trustworthy guidance with citations.
- Demonstrate feasibility and outline pathways for workflow integration and future expansion.

Question Answering using LLM - Model Name and Parameters

- Model = gpt-3.5-turbo
- Max Tokens = 512
- Temperature = 0
- Nucleus Sampling = Top_P = 0.95

Question Answering using LLM - Question 1: What is the protocol for managing sepsis in a critical care unit?

The protocol for managing sepsis in a critical care unit typically involves the following steps:

1. Early recognition and diagnosis: Prompt recognition of sepsis is crucial for initiating timely treatment. This involves assessing the patient's clinical signs and symptoms, as well as conducting laboratory tests to confirm the presence of infection.
2. Source control: Identifying and treating the source of infection is essential for managing sepsis. This may involve surgical intervention to drain abscesses or remove infected tissue.
3. Antibiotic therapy: Broad-spectrum antibiotics are typically initiated as soon as possible to target the suspected or identified pathogen. Antibiotic therapy may be adjusted based on culture results and the patient's response to treatment.
4. Fluid resuscitation: Patients with sepsis often experience hypotension and hypoperfusion, requiring aggressive fluid resuscitation to restore adequate tissue perfusion. Monitoring of fluid balance and hemodynamic parameters is essential to prevent fluid overload.
5. Vasopressor therapy: In cases of septic shock, vasopressor medications may be necessary to maintain adequate blood pressure and perfusion to vital organs. Close monitoring of hemodynamic parameters is essential to guide vasopressor therapy.
6. Supportive care: Patients with sepsis may require supportive care, such as mechanical ventilation, renal replacement therapy, and nutritional support. Close monitoring of organ function and vital signs is essential to guide treatment decisions.
7. Continuous monitoring and reassessment: Continuous monitoring of the patient's clinical status, laboratory values, and response to treatment is essential for guiding ongoing management. Reassessment of the patient's condition and treatment plan should be conducted regularly to ensure optimal care.

Overall, the management of sepsis in a critical care unit requires a multidisciplinary approach involving intensive care physicians, infectious disease specialists, nurses, and other healthcare providers. Timely recognition, aggressive treatment, and close monitoring are essential for improving outcomes in patients with sepsis.

Question Answering using LLM - 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?

Common symptoms of appendicitis include:

1. Sudden pain that starts around the navel and moves to the lower right abdomen
2. Nausea and vomiting
3. Loss of appetite
4. Fever
5. Constipation or diarrhea
6. Swelling or tenderness in the abdomen

Appendicitis cannot be cured with medicine alone. The standard treatment for appendicitis is surgical removal of the appendix, known as an appendectomy. This procedure is typically done as soon as possible after the diagnosis to prevent the appendix from rupturing and causing a serious infection in the abdomen. In some cases, a laparoscopic appendectomy may be performed, which is a minimally invasive procedure that involves making small incisions in the abdomen to remove the appendix. In more severe cases, an open appendectomy may be necessary, which involves making a larger incision in the abdomen to remove the appendix.

Question Answering using LLM - 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?

Sudden patchy hair loss, also known as alopecia areata, can be a distressing condition for those affected. The exact cause of alopecia areata is not fully understood, but it is believed to be an autoimmune disorder where the immune system mistakenly attacks hair follicles, leading to hair loss.

Some effective treatments and solutions for addressing sudden patchy hair loss include:

1. Corticosteroid injections: Injecting corticosteroids directly into the bald patches can help reduce inflammation and stimulate hair growth.
2. Topical corticosteroids: Applying corticosteroid creams or ointments to the affected areas can help reduce inflammation and promote hair regrowth.
3. Topical minoxidil: Minoxidil is a medication that is commonly used to treat male and female pattern baldness. It can also be effective in promoting hair regrowth in cases of alopecia areata.
4. Anthralin cream: Anthralin is a medication that is applied to the scalp to help stimulate hair growth in cases of alopecia areata.
5. Immunotherapy: This treatment involves applying chemicals like diphencyprone or squaric acid dibutyl ester to the scalp to provoke an allergic reaction and stimulate hair regrowth.
6. Hair transplant: In severe cases of alopecia areata, hair transplant surgery may be an option to restore hair growth in the affected areas.

It is important to consult with a dermatologist or healthcare provider to determine the best treatment plan for addressing sudden patchy hair loss. Additionally, managing stress, maintaining a healthy diet, and avoiding harsh hair treatments can help improve the condition of the scalp and promote hair growth.

Question Answering using LLM - 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?

The recommended treatments for a person who has sustained a physical injury to brain tissue will depend on the severity and specific nature of the injury. Some common treatments may include:

1. Rest and monitoring: In cases of mild traumatic brain injury (concussion), rest and close monitoring of symptoms may be all that is needed for recovery.
2. Medications: Depending on the symptoms and severity of the injury, medications may be prescribed to manage pain, reduce inflammation, prevent seizures, or improve cognitive function.
3. Rehabilitation therapy: Physical, occupational, and speech therapy may be recommended to help the individual regain lost skills, improve coordination and balance, and enhance communication abilities.
4. Cognitive therapy: Cognitive rehabilitation therapy may be used to help improve memory, attention, problem-solving skills, and other cognitive functions that may be affected by the brain injury.
5. Assistive devices: Depending on the nature of the impairment, assistive devices such as wheelchairs, walkers, communication aids, or other adaptive equipment may be recommended to help the individual function more independently.
6. Surgery: In cases of severe brain injury, surgery may be necessary to remove blood clots, repair damaged tissue, or relieve pressure on the brain.
7. Counseling and support: Emotional and psychological support may be needed to help the individual and their family cope with the challenges of living with a brain injury.

It is important for individuals who have sustained a brain injury to work closely with a healthcare team, including doctors, therapists, and other specialists, to develop a comprehensive treatment plan tailored to their specific needs and goals.

Question Answering using LLM - Overall Observations

- **Structure:** The DataFrame is well-structured with two columns: questions and base_prompt_responses. This format is suitable for direct comparison and further evaluation.
- **Question Content:** The questions cover diverse medical domains (critical care, general surgery, dermatology, neurology), indicating a need for a comprehensive knowledge base.
- **Response Characteristics:** The base_prompt_responses are generally well-articulated and directly address the questions. However, without external context, their factual accuracy or completeness cannot be fully ascertained from the DataFrame alone. They represent the LLM's inherent knowledge without any external retrieval mechanism.
- **Baseline for Comparison:** This DataFrame serves as an important baseline. It allows us to compare the effectiveness of subsequent methods (like prompt engineering and RAG) against the model's performance with a simple, direct prompt.

Question Answering using LLM with Prompt Engineering - System Prompt

You are MedAssist, an AI medical assistant for licensed healthcare professionals.

Your role:

- Provide concise, evidence-based medical information from authoritative sources.
- Support clinical reasoning by summarizing data, outlining differentials, describing mechanisms, and identifying red flags.
- Improve efficiency by presenting information in clear, structured formats (lists, tables, brief summaries).

Requirements:

- Do NOT provide definitive diagnoses, prescribe treatments, or give patient-directed advice. Always defer to clinician judgment.
- Do NOT fabricate information, citations, or guidelines. State when evidence is uncertain or limited.
- Prioritize accuracy, relevance, and up-to-date medical knowledge.
- Communicate in a professional, direct, and technical tone suitable for clinicians.

Question Answering using LLM with Prompt Engineering- Question 1 - What is the protocol for managing sepsis in a critical care unit?

Managing sepsis in a critical care unit involves prompt recognition, aggressive resuscitation, and targeted treatment to improve outcomes. The protocol typically follows guidelines such as the Surviving Sepsis Campaign. Here is a summary of key components in managing sepsis:

1. Recognition and Initial Assessment

- Early recognition of sepsis using clinical criteria (e.g., SIRS criteria, qSOFA score).
- Prompt assessment of hemodynamic status, perfusion, and organ dysfunction.

2. Resuscitation

- Initiate early goal-directed therapy with a focus on hemodynamic stabilization.
- Administer intravenous fluids judiciously to optimize perfusion.
- Initiate vasopressors if needed to maintain adequate mean arterial pressure.

3. Antimicrobial Therapy

- Administer broad-spectrum antibiotics promptly within the first hour of recognition.
- Adjust antibiotics based on culture results and de-escalate therapy as appropriate.

4. Source Control

- Identify and control the source of infection through procedures like drainage or debridement.
- Consider surgical intervention if necessary.

Question Answering using LLM with Prompt Engineering- Question 1 - What is the protocol for managing sepsis in a critical care unit? (Cont'd)

Managing sepsis in a critical care unit involves prompt recognition, aggressive resuscitation, and targeted treatment to improve outcomes. The protocol typically follows guidelines such as the Surviving Sepsis Campaign. Here is a summary of key components in managing sepsis:

5. Supportive Care

- Provide respiratory support with mechanical ventilation if needed.
- Monitor and manage electrolyte imbalances, coagulopathy, and other organ dysfunctions.

6. Monitoring and Reassessment

- Continuously monitor vital signs, laboratory parameters, and response to treatment.
- Reassess the patient's condition regularly to guide further management.

7. Consideration of Adjunctive Therapies

- Consider adjunctive therapies like corticosteroids, activated protein C (if indicated), and immunomodulatory agents in specific cases.

8. Early Involvement of Multidisciplinary Team

- Involve infectious disease specialists, pharmacists, and other healthcare professionals early in the management of sepsis.

9. Family Communication and Support

- Provide clear communication with the patient's family regarding the diagnosis, treatment plan, and prognosis.
- Offer emotional support and involve palliative care services if needed.

Remember, individual patient factors and local guidelines may influence the specific management approach. Regular reassessment and adjustment of therapy are crucial in the dynamic management of sepsis in the critical care setting.

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?

Sudden patchy hair loss, known as alopecia areata, presents as well-defined bald spots on the scalp. The exact cause is not fully understood but is believed to involve an autoimmune reaction targeting hair follicles. Here are some key points to consider:

Possible Causes:

1. **Autoimmune Reaction:** Immune system mistakenly attacks hair follicles.
2. **Genetics:** Family history may play a role.
3. **Stress:** Emotional or physical stress can trigger episodes.
4. **Environmental Factors:** Allergens, infections, or toxins may contribute.

Differential Diagnoses:

1. **Tinea Capitis:** Fungal infection causing patchy hair loss.
2. **Trichotillomania:** Compulsive hair pulling leading to bald spots.
3. **Secondary Syphilis:** Can manifest as patchy hair loss.

Treatment Options:

1. **Topical Corticosteroids:** Reduce inflammation and promote hair regrowth.
2. **Intralesional Corticosteroid Injections:** Directly into affected areas for more severe cases.
3. **Topical Immunotherapy:** Provokes an allergic reaction to stimulate hair growth.
4. **Minoxidil:** Over-the-counter option to promote hair regrowth.
5. **Platelet-Rich Plasma (PRP) Therapy:** Stimulates hair follicles.

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? (Cont'd)

Monitoring and Referral:

1. **Regular Follow-up:** Monitor progression and response to treatment.
2. **Dermatologist Referral:** For severe cases or if initial treatments are ineffective.
3. **Psychological Support:** Address emotional impact of hair loss.

Red Flags:

- **Rapid Spread:** Sudden extensive hair loss may indicate a more serious condition.
- **Scalp Irritation:** Itching, pain, or redness accompanying hair loss.
- **Systemic Symptoms:** Fever, weight loss, or fatigue.

It's crucial to conduct a thorough evaluation to confirm the diagnosis and tailor treatment accordingly. Referral to a dermatologist can provide specialized care and management strategies for alopecia areata.

Question Answering using LLM with Prompt Engineering- Question 3 - 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?

Sudden patchy hair loss, known as alopecia areata, presents as well-defined bald spots on the scalp. The exact cause is not fully understood but is believed to involve an autoimmune reaction targeting hair follicles. Here are some key points to consider:

Possible Causes:

1. **Autoimmune Reaction:** Immune system mistakenly attacks hair follicles.
2. **Genetics:** Family history may play a role.
3. **Stress:** Emotional or physical stress can trigger episodes.
4. **Environmental Factors:** Allergens, infections, or toxins may contribute.

Differential Diagnoses:

1. **Tinea Capitis:** Fungal infection causing patchy hair loss.
2. **Trichotillomania:** Compulsive hair pulling leading to bald spots.
3. **Secondary Syphilis:** Can manifest as patchy hair loss.

Treatment Options:

1. **Topical Corticosteroids:** Reduce inflammation and promote hair regrowth.
2. **Intralesional Corticosteroid Injections:** Directly into affected areas for more severe cases.
3. **Topical Immunotherapy:** Provokes an allergic reaction to stimulate hair growth.
4. **Minoxidil:** Over-the-counter option to promote hair regrowth.
5. **Platelet-Rich Plasma (PRP) Therapy:** Stimulates hair follicles.

Question Answering using LLM with Prompt Engineering- Question 3 - 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? (Cont'd)

Monitoring and Referral:

1. **Regular Follow-up:** Monitor progression and response to treatment.
2. **Dermatologist Referral:** For severe cases or if initial treatments are ineffective.
3. **Psychological Support:** Address emotional impact of hair loss.

Red Flags:

- **Rapid Spread:** Sudden extensive hair loss may indicate a more serious condition.
- **Scalp Irritation:** Itching, pain, or redness accompanying hair loss.
- **Systemic Symptoms:** Fever, weight loss, or fatigue.

It's crucial to conduct a thorough evaluation to confirm the diagnosis and tailor treatment accordingly.

Referral to a dermatologist can provide specialized care and management strategies for alopecia areata.

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?

Treatment for brain injuries depends on the severity and type of injury. Here are some general approaches commonly used in managing brain injuries:

1. **Acute Management**:

- **Stabilization**: Ensure airway, breathing, and circulation are maintained.
- **Monitoring**: Monitor intracranial pressure, neurological status, and vital signs.
- **Imaging**: CT or MRI scans to assess the extent of the injury.

2. **Medical Interventions**:

- **Surgery**: May be necessary to remove hematomas, repair skull fractures, or relieve pressure on the brain.
- **Medications**:
 - **Anti-seizure drugs**: To prevent seizures.
 - **Diuretics**: To reduce brain swelling.
 - **Pain medications**: To manage discomfort.
 - **Stimulants or antidepressants**: To address cognitive or emotional symptoms.

3. **Rehabilitation**:

- **Physical therapy**: To improve motor skills and coordination.
- **Occupational therapy**: To regain daily living skills.
- **Speech therapy**: To address communication difficulties.
- **Cognitive therapy**: To improve memory, attention, and problem-solving skills.

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? (Cont'd)

4. **Psychological Support**:

- **Counseling**: To address emotional and psychological challenges.
- **Support groups**: To connect with others facing similar challenges.

5. **Lifestyle Modifications**:

- **Healthy diet**: To support brain health and overall recovery.
- **Regular exercise**: To improve physical and mental well-being.
- **Adequate rest**: To promote healing and recovery.

It's crucial for individuals with brain injuries to receive individualized care from a multidisciplinary team including neurologists, neurosurgeons, physical therapists, occupational therapists, speech therapists, and psychologists. The treatment plan should be tailored to the specific needs and goals of the patient, considering the nature and extent of the brain injury.

Question Answering using LLM with Prompt Engineering - Overall Observations

Category	Observation
Factual Accuracy	Improves structure and reduces some hallucinations, but cannot guarantee clinically accurate information without external grounding.
Behavior Control	Prompts can guide tone, format, and reasoning steps, but cannot force the model to avoid incorrect content .
Consistency	More predictable than an untuned LLM, yet still prone to variability depending on input phrasing.
Handling Out-of-Scope Questions	Better at deflecting when instructed, but still may produce confident-sounding incorrect answers .
Medical Specificity	Able to mimic medical language, but lacks protocol-specific details unless they exist in its pretrained knowledge.
Adaptability to Local Protocols	Limited; cannot reliably incorporate institution-specific guidelines .
Operational Overhead	Low effort to iterate on prompts, but prompts alone cannot correct factual errors or fill knowledge gaps.
Best Use Cases	Structuring responses, generating summaries, improving readability—not for verified clinical decision support .

Data Preparation for RAG

- Dataset used= Merck Manuals.pdf
- Parameters used to split the data (*chunk_overlap, chunk_size*)
 - Chunk_overlap=200
 - chunk_size=100
- RAG Parameters
 - k=5
 - Max tokens =128
 - Temp = 0
 - Top p=0.95

Question Answering using RAG

- System and user prompt
 - As an AI medical assistant, your primary role is to provide accurate, relevant, and comprehensive answers to medical queries based solely on the provided context. Do not use any outside knowledge. If the answer cannot be found within the given context, state that clearly and politely.
 - Based on the following context, please answer the question. If the answer is not available in the context, please state "I am sorry, but I do not have information about this topic."

Question Answering using RAG - Question #1

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

- The protocol for managing sepsis in a critical care unit includes the following steps:
 - 1. Prompt empiric therapy with parenteral antibiotics based on educated guesses of the suspected source, clinical setting, and knowledge of causative organisms.
 - 2. Consideration of antibiotic regimens such as gentamicin or tobramycin with a 3rd-generation cephalosporin or ceftazidime plus a fluoroquinolone for septic shock.
 - 3. Addition of vancomycin if resistant staphylococci or enterococci are suspected.
 - 4. Inclusion of a drug effective against anaerob

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?

The common symptoms of appendicitis include epigastric or periumbilical pain followed by nausea, vomiting, anorexia, and then shifting to the right lower quadrant. Pain increases with cough and motion. Classic signs include right lower quadrant direct and rebound tenderness at McBurney's point. Additional signs may include Rovsing sign, psoas sign, or obturator sign. Low-grade fever is also common. However, it is important to note that appendicitis is primarily treated through surgical removal of the inflamed appendix, known as an appendectomy. Antibiotics are used before surgery to reduce the risk of

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?

I am sorry, but I do not have information about this topic.

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?

Initial treatment for a person who has sustained a physical injury to brain tissue, resulting in temporary or permanent impairment of brain function includes ensuring a reliable airway, maintaining adequate ventilation, oxygenation, and blood pressure. Surgery may be needed in patients with more severe injuries to monitor and treat intracranial pressure, decompress the brain if intracranial pressure is increased, or remove intracranial hematomas. Additionally, in the first few days after the injury, it is important to maintain adequate brain perfusion and oxygenation and prevent complications of altered sensorium. Subsequently, many patients may require rehabilitation.

Overall Comments and Observations for Question Answering using RAG

1. RAG Provides Clinically Grounded Answers

- Every response is anchored to verified medical guidelines, reducing hallucinations.
- Ensures accuracy for dosages, contraindications, prognosis criteria, and triage steps.
- Safer than LLM-only or prompt-engineered LLMs, which rely on general knowledge.

2. Better Handling of Out-of-Scope Questions

- RAG explicitly states when medical information is not available in the source documents.
- Prevents unsafe speculation—a common behavior in standard LLMs.

3. Richer and More Precise Medical Detail

- Retrieves specific language directly from medical manuals.
- Enables protocol-consistent responses: red flags, treatment steps, clinical thresholds.
- LLMs without retrieval generalize and may misstate critical details.

Overall Comments and Observations for Question Answering using RAG (Cont'd)

4. Retrieval Tuning Improves Performance

- k = number of chunks/documents retrieved
- Optimal medical k (typically 3–5) ensures:
 - Enough clinical context for accurate answers
 - Not so much that the model becomes noisy or distracted
- This tuning capability doesn't exist in LLM-only systems.

LLM-only and prompt-engineered models improve usability but cannot meet the factual accuracy and safety standards required for clinical environments.

5. Operational Advantages

- Updating medical guidelines automatically improves the system—no retraining required.
- Easy to customize for hospital-specific protocols and terminology.
- Supports traceable, auditable outputs required for clinical governance.

Executive Takeaway

RAG is the only approach that reliably delivers safe, accurate, and protocol-aligned medical answers.

LLM-only and prompt-engineered models improve usability but cannot meet the factual accuracy and safety standards required for clinical environments.

Output Evaluation - Evaluation prompt for Groundedness

You are an expert at evaluating the 'groundedness' of a generated answer. Groundedness refers to whether the answer is fully supported by the provided context. Your task is to analyze the given question, context, and a generated answer. Determine if the answer contains any information that is not explicitly present or inferable from the context. If the answer introduces new facts or makes statements that cannot be traced back to the context, it is not fully grounded.

Rate the groundedness on a scale from 1 to 5, where:

- 1: The answer is entirely ungrounded; most or all information comes from outside the context.
- 2: The answer contains significant ungrounded information, with only minor parts supported by the context.
- 3: The answer is partially grounded; some parts are directly supported, but other significant parts are not.
- 4: The answer is mostly grounded; only minor details or interpretations are not directly supported by the context.
- 5: The answer is fully grounded; every piece of information in the answer can be directly found or logically inferred from the provided context.

Provide only the numerical rating as your response.

Output Evaluation - Evaluation prompt for Relevance

You are an expert at evaluating the 'relevance' of a generated answer. Relevance refers to how well the answer addresses the user's question and if it directly answers the core intent of the question. Your task is to analyze the given question, context, and a generated answer. Determine if the answer is focused on the question, avoids extraneous information, and provides a direct response to the query.

Rate the relevance on a scale from 1 to 5, where:

- 1: The answer is entirely irrelevant and does not address the question at all.
- 2: The answer is mostly irrelevant; it touches on the topic but does not directly answer the question.
- 3: The answer is partially relevant; it addresses some aspects of the question but misses key points or includes extraneous information.
- 4: The answer is mostly relevant; it addresses the core question but might include some minor irrelevant details or lack completeness.
- 5: The answer is fully relevant; it directly and comprehensively answers the question without including irrelevant information.

Provide only the numerical rating as your response.

Output Evaluation - Model Scores Groundness & Relevance - Base Prompt

- Question 1
 - Groundedness Score - 4
 - Relevance Score - 4
- Question 2
 - Groundedness Score - 4
 - Relevance Score - 4
- Question 3
 - Groundedness Score - 4
 - Relevance Score - 4
- Question 4
 - Groundedness Score - 4
 - Relevance Score - 4

Output Evaluation - Model Scores Groundness & Relevance - Prompt Engineering

- Question 1
 - Groundedness Score - 4
 - Relevance Score - 4
- Question 2
 - Groundedness Score - 5
 - Relevance Score - 5
- Question 3
 - Groundedness Score - 4
 - Relevance Score - 5
- Question 4
 - Groundedness Score - 4
 - Relevance Score - 4

Output Evaluation - Model Scores Groundness & Relevance - RAG

- Question 1
 - Groundedness Score - 4
 - Relevance Score - 4
- Question 2
 - Groundedness Score - 4
 - Relevance Score - 4
- Question 3
 - Groundedness Score - 4
 - Relevance Score - 4
- Question 4
 - Groundedness Score - 4
 - Relevance Score - 4

Output Evaluation - Overall Comments and Observations

1. Base Prompt Responses:

- **Groundedness:** These responses are generated solely from the LLM's internal knowledge without any specific external context provided at the generation stage. When evaluated against *retrieved* context (as our generate_ground_relevance_response function implicitly does), they might score moderately (e.g., 3-4). The LLM might generate factually correct information, but it may not directly align with the specific phrases or details available in the retrieved context for the evaluation.
- **Relevance:** Generally, base prompt responses are expected to be relevant (e.g., 4-5) as the LLM attempts to directly answer the question. However, without further guidance, they might include general information or omit critical details.

2. Prompt Engineering Responses:

- **Groundedness:** With a well-crafted system prompt guiding the AI to be accurate and comprehensive, the groundedness might improve slightly compared to base prompts (e.g., 4-5). The prompt encourages the model to stick to facts and explain concepts clearly, which can indirectly lead to better alignment with potentially retrieved context during evaluation. However, the model still primarily relies on its training data at this stage.
- **Relevance:** The prompt-engineered responses should demonstrate high relevance (e.g., 4-5) due to the explicit instructions to explain symptoms, causes, treatments, and protocols in a precise manner.

3. RAG Responses:

- **Groundedness:** This is where RAG is expected to shine. By explicitly retrieving relevant document chunks and instructing the LLM to answer *only* based on this context, RAG responses should achieve the highest groundedness scores (e.g., 4-5). This minimizes hallucination and ensures traceability to source material.
- **Relevance:** RAG responses are typically highly relevant (e.g., 4-5) because the retrieved context is directly pertinent to the user's query, and the LLM is tasked with synthesizing an answer from this specific information.
- **Handling of Unknown Information:** A key advantage of RAG is its ability to explicitly state when it cannot find information within the provided context, as seen with Question 3, rather than generating potentially incorrect or generalized answers.

APPENDIX

Data Background and Contents

- Dataset for RAG
 - Merck Manuals.pdf
 - The **Merck Manuals** are medical references published by the American pharmaceutical company Merck & Co., that cover a wide range of medical topics, including disorders, tests, diagnoses, and drugs. The manuals have been published since 1899, when Merck & Co. was still a subsidiary of the German company Merck.
 - The manual is provided as a PDF with over 4,000 pages divided into 23 section
- Model Used for LLM Comparisons
 - gpt-3.5-turbo
 - It is optimized for chat and other natural language tasks, including understanding and generating text and code



Power Ahead!

