



# Medical Assistant Business Applications with LLMs

Leveraging Large Language Models and Retrieval-Augmented Generation to  
Transform Clinical Decision Support at the Point of Care

MARCH 29, 2026

# Contents & Agenda

This presentation walks through a structured exploration of LLM capabilities applied to clinical decision support. From raw model behavior to production-ready RAG deployment and rigorous output evaluation.

01

---

## Question Answering Using LLM

Baseline evaluation of raw GPT-4o performance on clinical queries without any system-level guidance

02

---

## Question Answering with Prompt Engineering

Applying structured system prompts — role definition, negative constraints, and output formatting — to shape model behavior

03

---

## Data Preparation for RAG

Ingesting and indexing the 4,104-page Merck Manual into a high-dimensional vector database for semantic retrieval

04

---

## Question Answering Using RAG

Combining optimized prompts with retrieved context chunks to deliver grounded, source-cited clinical responses

05

---

## Output Evaluation

Scoring every response using an "LLM-as-a-Judge" framework across Groundedness and Relevance dimensions

# Executive Summary: The Problem

## CLINICAL INFORMATION OVERLOAD

The healthcare industry faces a systemic crisis: the sheer volume of medical research and diagnostic data, exemplified by the **4,000+ page Merck Manual**, far exceeds the cognitive processing capacity of clinicians operating in high-pressure environments. This shows up as a **retrieval and synthesis gap**.

### Latency in Decision-Making

Manually sifting through extensive clinical manuals during time-sensitive emergencies leads to delayed diagnoses, treatment lag, and avoidable adverse outcomes

### Lack of Standardized Retrieval

Relying on individual memory or fragmented search methods results in inconsistent application of clinical protocols

### Accuracy vs. Speed Trade-off

The pressure to deliver rapid care directly conflicts with the need to cross-reference the most current, trusted medical evidence, forcing clinicians to choose between speed and safety

# The Solution

- 📄 An **Automated Clinical Decision Support System** built on Retrieval-Augmented Generation (RAG) — designed to close the gap between vast medical knowledge and real-time clinical application.

This system must deliver grounded, evidence-based interventions with **100% source transparency**, ensuring every recommendation is traceable to an approved medical authority.

---

## Core Design Principles



### Centralized Knowledge Access

Streamline retrieval from a single, authoritative vector-indexed medical corpus — no fragmented searches, no memory reliance



### Grounded, Cited Responses

Every answer surfaces a specific source citation (e.g., "Source: Page 2456"), making AI output verifiable and clinically trustworthy



### Improved Patient Outcomes

Faster, more consistent decision support directly reduces diagnostic latency and supports safer, protocol-aligned treatment paths

# Recommendations

## 1. Implement "Verified-Only" Clinical Dashboards

**Finding:** The RAG system delivers grounded answers with "Source: Page XXX" citations, while the base LLM frequently hallucinates general medical knowledge without verification

**Suggestion:** Replace general-purpose AI chat interfaces with a **Source-Locked Knowledge Portal**. Configure the UI to display answers only when the **Groundedness Score  $\geq 4.5$**

**Impact:** Reduces automated medical errors and increases physician trust by linking directly to hospital-approved manuals (Merck)

## 2. Automated Triage Support

**Finding:** The optimized prompt prioritized "1-hour bundles" and "Initial Resuscitation" protocols at the top of every clinical response — exactly what a triage nurse needs first

**Suggestion:** Integrate the RAG engine into the **Electronic Health Record (EHR)** alert system. When a patient's vitals trigger a sepsis alert, the system auto-pulls the top 3 treatment steps

**Impact:** Filters 4,000 pages of manual data into 3 actionable steps — potentially saving lives within critical care windows

# Recommendations (3 & 4)

## 3. Scale Staff Training with "AI-as-a-Coach"

**Finding:** The "LLM-as-a-Judge" evaluation function objectively grades the quality and groundedness of medical information on a 1–5 scale

**Suggestion:** Deploy the evaluation framework as an **Automated Medical Residency Auditor**, using Judge prompts to review resident notes and treatment plans against the Merck Manual

**Impact:** Provides instant, 24/7 feedback to junior staff on protocol alignment, reducing the supervisory burden on senior clinicians and accelerating clinical competency development

## 4. Operational Cost Savings via Document Digitization

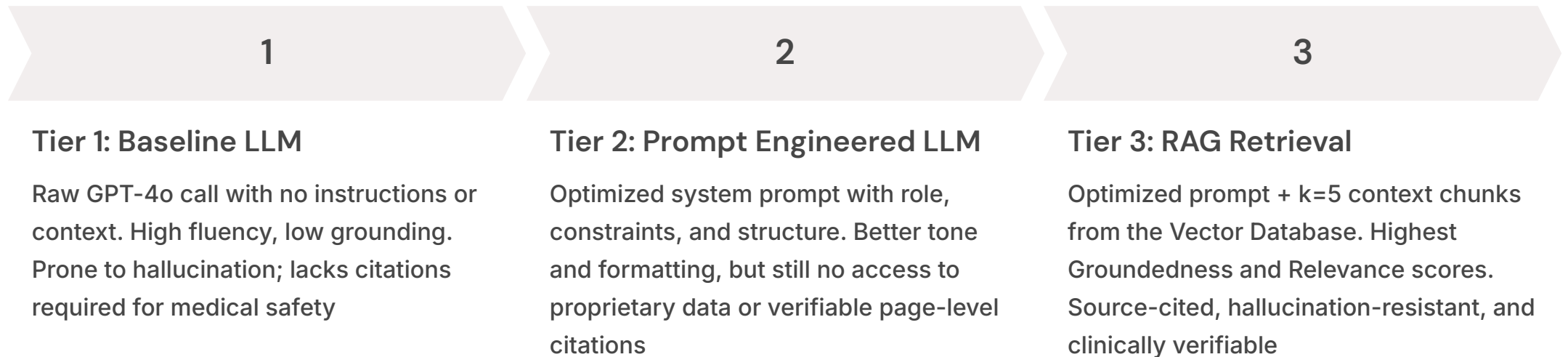
**Finding:** 4,104 pages were successfully indexed into a searchable vector database that costs pennies per query, a fraction of the cost of physical manual subscriptions or manual search time.

**Suggestion:** Expand the RAG system to include internal hospital **SOPs** and insurance billing codes within a **Centralized Proprietary Knowledge Base**

**Impact:** Eliminates an estimated **15–20% of a physician's day** lost to administrative search time, while ensuring every department operates from a single, unified version of the truth

# Business Problem & Solution: Three-Tiered Approach

To evaluate the impact of AI-assisted clinical decision support, we structured our methodology as a controlled three-tier experiment — moving from raw model output to fully grounded RAG retrieval



**Key Insight:** Only Tier 3 (RAG) solved the Information Overload problem — filtering 4,000 pages into 5 targeted segments, delivering both speed and safety at the point of care

# RAG Methodology: How It Works

## The Core Architecture

Our RAG prototype moves beyond general-purpose AI chatting to deliver **Source-Grounded Decision Support**, built on the full 4,104-page Merck Manual, indexed into a ChromaDB vector database with **12,492 granular text chunks**.

Every response is synthesized exclusively from retrieved manual content. The model is explicitly **forbidden from using external training knowledge** — ensuring zero hallucination risk for clinical outputs.

1

### Data Ingestion & Indexing

4,104-page Merck Manual converted into a high-dimensional ChromaDB vector database, segmented into 12,492 chunks for granular retrieval

2

### Contextual Retrieval (k=5)

Semantic search identifies the five most relevant manual segments per query, capturing multi-step protocols (e.g., sepsis management) across distributed pages

3

### Grounded Synthesis

Retrieved context is passed to GPT-4o with a strict system prompt. Output is grounded solely in manual text, complete with page-level citations

4

### Multi-Dimensional Evaluation

"LLM-as-a-Judge" framework grades every response 1–5 on **Groundedness** (zero hallucinations) and **Relevance** (clinical utility and directness)



# Question Answering using LLM

A comparative study of **Base LLM**, **Prompt Engineering**, and **Retrieval-Augmented Generation (RAG)** for clinical decision support — evaluated on groundedness, relevance, and medical safety using the Merck Manual as the authoritative source.

MODEL: GPT-4O

TEMPERATURE: 0

MAX TOKENS: 1000

EXPERIMENTAL

# Experimental Setup

All testing tiers used identical parameters to ensure a scientifically controlled environment.

---

## Constant Parameters

- **Model:** gpt-4o (OpenAI)
- **Temperature:** 0 — absolute determinism for clinical safety
- **Max Tokens:** 1000 — runway for multi-step protocols
- **Top\_P:** 1.0 — standard for greedy decoding

## 4 Test Questions

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



## Base LLM Responses

The baseline GPT-4o model was tested without any system prompt or retrieval augmentation. While responses demonstrated strong medical intuition, they revealed a critical "**Verification Gap**" - the inability to cite sources, risking clinical safety

# Base LLM Responses – Question 1

Managing sepsis in a critical care unit involves a systematic approach to stabilize the patient, identify and treat the source of infection, and support organ function. The protocol generally follows guidelines such as those from the Surviving Sepsis Campaign. Here is an overview of the typical steps involved:

- 1. Early Recognition and Diagnosis:** - Monitor for signs of sepsis, including fever, increased heart rate, increased respiratory rate, altered mental status, and laboratory markers like elevated lactate levels. - Use screening tools and protocols to identify sepsis early.
- 2. Initial Resuscitation:** - Begin treatment immediately, ideally within the first hour of recognition. - Administer intravenous fluids (usually crystalloids) to maintain adequate blood pressure and tissue perfusion. - Target a mean arterial pressure (MAP) of 65 mmHg or higher.
- 3. Antibiotic Therapy:** - Administer broad-spectrum antibiotics as soon as possible, ideally within the first hour. - Adjust antibiotic therapy based on culture results and sensitivity testing.
- 4. Source Control:** - Identify and control the source of infection, which may involve surgical intervention, drainage of abscesses, or removal of infected devices.
- 5. Hemodynamic Support:** - Use vasopressors, such as norepinephrine, if fluid resuscitation is insufficient to maintain MAP. - Consider inotropic agents if cardiac output is low despite adequate fluid resuscitation.
- 6. Monitoring and Supportive Care:** - Continuously monitor vital signs, urine output, and laboratory parameters. - Provide supplemental oxygen or mechanical ventilation if needed to maintain adequate oxygenation. - Monitor and support organ function, including renal replacement therapy for kidney failure if necessary.
- 7. Adjunctive Therapies:** - Consider corticosteroids for patients with septic shock unresponsive to fluid and vasopressor therapy. - Manage blood glucose levels, aiming for a target of less than 180 mg/dL.
- 8. Reassessment and Adjustment:** - Regularly reassess the patient's condition and adjust treatment plans accordingly. - De-escalate antibiotics based on culture results and clinical improvement.
- 9. Communication and Documentation:** - Maintain clear communication with the healthcare team and document all interventions and patient responses.
- 10. Education and Prevention:** - Educate healthcare staff and patients about sepsis recognition and prevention strategies.

This protocol may vary based on institutional guidelines and the specific needs of the patient. It is crucial to follow the latest evidence-based guidelines and collaborate with a multidisciplinary team for optimal patient outcomes.

# Base LLM Responses – Question 2

Appendicitis is an inflammation of the appendix and is considered a medical emergency. Common symptoms include: 1. **Abdominal Pain:** Typically starts near the navel and then shifts to the lower right abdomen. The pain usually becomes sharper and more severe over time. 2. **Loss of Appetite:** A sudden loss of appetite is common. 3. **Nausea and Vomiting:** Often follows the onset of abdominal pain. 4. **Fever:** A low-grade fever may be present. 5. **Constipation or Diarrhea:** Some people experience changes in bowel habits. 6. **Abdominal Swelling:** The abdomen may become swollen and tender to touch. 7. **Inability to Pass Gas:** This can be a sign of a blockage. Appendicitis is typically not cured with medication alone. Antibiotics may be administered to manage infection, but they are usually not sufficient as a standalone treatment. The standard treatment for appendicitis is surgical removal of the appendix, known as an **appendectomy**. There are two main types of appendectomy: 1. **Open Appendectomy:** Involves a single larger incision in the lower right abdomen to remove the appendix. This method is often used if the appendix has ruptured or if there is an abscess. 2. **Laparoscopic Appendectomy:** Involves several small incisions and the use of a camera and special instruments to remove the appendix. This method is less invasive, typically results in a quicker recovery, and is preferred when possible. Prompt surgical intervention is crucial to prevent complications such as a ruptured appendix, which can lead to peritonitis or abscess formation. If you suspect appendicitis, it is important to seek medical attention immediately.

# Base LLM Responses – Question 3

Sudden patchy hair loss, often presenting as localized bald spots on the scalp, is commonly associated with a condition called alopecia areata. This autoimmune disorder causes the body's immune system to mistakenly attack hair follicles, leading to hair loss. While the exact cause of alopecia areata is not fully understood, it is believed to involve a combination of genetic predisposition and environmental factors. Stress and other autoimmune conditions may also play a role in triggering or exacerbating the condition. Effective treatments and solutions for addressing alopecia areata include:

- 1. Topical Corticosteroids:** These are anti-inflammatory medications applied directly to the bald spots to help suppress the immune response and promote hair regrowth.
- 2. Intralesional Corticosteroid Injections:** For more localized and severe patches, corticosteroids can be injected directly into the affected areas to stimulate hair growth.
- 3. Topical Immunotherapy:** This involves applying chemicals like diphencyprone (DPCP) to the scalp to provoke an allergic reaction, which can alter the immune response and encourage hair regrowth.
- 4. Minoxidil (Rogaine):** An over-the-counter topical treatment that can help stimulate hair growth in some cases of alopecia areata.
- 5. Anthralin:** A topical medication that alters immune function in the skin and can be used to treat patchy hair loss.
- 6. Oral Corticosteroids:** In more severe cases, systemic corticosteroids may be prescribed to reduce inflammation and immune activity throughout the body.
- 7. JAK Inhibitors:** These are newer oral medications that target specific pathways in the immune system and have shown promise in treating alopecia areata.
- 8. Platelet-Rich Plasma (PRP) Therapy:** This involves injecting concentrated platelets from the patient's own blood into the scalp to promote hair growth.
- 9. Lifestyle and Dietary Changes:** Managing stress, maintaining a balanced diet, and ensuring adequate intake of vitamins and minerals like biotin, zinc, and iron can support overall hair health.
- 10. Supportive Therapies:** Psychological support or counseling can be beneficial, as hair loss can have significant emotional and psychological impacts. It's important for individuals experiencing sudden patchy hair loss to consult with a dermatologist or healthcare provider to determine the underlying cause and appropriate treatment plan.

# Base LLM Responses – Question 4

Treating a brain injury, whether temporary or permanent, involves a multidisciplinary approach tailored to the individual's specific needs and the severity of the injury. Here are some common treatments and interventions:

- 1. Emergency Care:** - Immediate medical attention is crucial to stabilize the patient, prevent further injury, and address life-threatening conditions. - This may include surgery to relieve pressure on the brain, repair skull fractures, or remove hematomas.
- 2. Medication:** - Medications may be prescribed to manage symptoms such as pain, seizures, spasticity, or to reduce swelling in the brain. - Examples include diuretics, anti-seizure drugs, and coma-inducing drugs if necessary.
- 3. Rehabilitation Therapy:**
  - Physical Therapy:** To improve motor skills, strength, and coordination.
  - Occupational Therapy:** To help the patient relearn daily activities and improve fine motor skills.
  - Speech and Language Therapy:** To address communication difficulties and swallowing issues.
  - Cognitive Rehabilitation:** To improve memory, attention, problem-solving, and other cognitive functions.
- 4. Psychological Support:** - Counseling or psychotherapy can help address emotional and behavioral changes. - Support groups for patients and families can provide emotional support and practical advice.
- 5. Lifestyle Modifications:** - Implementing a healthy diet, regular exercise, and adequate sleep can support recovery. - Avoiding alcohol and drugs, which can impede recovery.
- 6. Assistive Devices:** - Depending on the level of impairment, devices such as wheelchairs, communication aids, or adaptive equipment for daily living may be necessary.
- 7. Long-term Monitoring and Follow-up:** - Regular follow-up appointments to monitor progress and adjust treatment plans as needed. - Ongoing assessment by neurologists, physiatrists, and other specialists.
- 8. Community and Social Support:** - Engaging with community resources and social services to assist with reintegration into daily life and work. The treatment plan should be individualized, taking into account the type and severity of the brain injury, the patient's overall health, and their personal and social circumstances. Collaboration among healthcare providers, patients, and families is essential for optimal recovery.

# Base LLM: Key Observations

## Logical Sequencing

The model consistently prioritized life-saving interventions first (e.g., "Emergency Care" for brain injury, "Initial Resuscitation" for sepsis), aligning with the goal of reducing **Information Overload**

## Diagnostic Accuracy vs. Generalization

Answers provided broad medical consensus but included "conversational filler" (e.g., *"It is crucial to follow the latest guidelines"*). These generalities will be replaced by specific Merck Manual citations in the RAG phase

## Safety Guardrails

The model correctly identified that appendicitis cannot be "cured" by medicine alone, preventing dangerous misinformation. This proves a strong internal "world model," but reinforces the need for RAG for specific dosages and surgical criteria

# Base LLM: Appendicitis & Hair Loss Responses

## Appendicitis (Q2)

Listed 7 common symptoms including abdominal pain migrating to the lower right, nausea, fever, and abdominal swelling. Correctly noted antibiotics are insufficient as standalone treatment.

- ❏ **Feedback:** Broad "knowledge-retrieval" task led to **Information Overload**. RAG can narrow to primary first-line treatments cited by medical authority.

## Patchy Hair Loss (Q3)

Identified alopecia areata as the likely cause — an autoimmune disorder attacking hair follicles. Listed treatments including topical corticosteroids, intralesional injections, and minoxidil.

- ❏ **Feedback:** High-quality summary but fails the **Medical Safety Test** — cannot prove its source. RAG forces citation of the exact manual page.



## CHAPTER 2

# Prompt Engineering

A specialized system prompt transformed the model into a **Clinical Decision Support AI**, enforcing structured reasoning, urgency-based prioritization, elimination of conversational filler, and a mandatory medical disclaimer on every response.

# The Clinical Reasoning System Prompt

You are a Clinical Decision Support AI. Your task is to provide high-fidelity, evidence-based medical information structured for a critical care environment.

## OPERATIONAL PROTOCOL:

1. **Clinical Reasoning:** Before providing the protocol, briefly identify the primary diagnostic criteria (e.g., SIRS, qSOFA, or SOFA scores).
2. **Prioritization:** List interventions in order of clinical urgency (e.g., the '1-hour bundle' for Sepsis).
3. **Negative Constraint:** Do not provide conversational filler, disclaimers at the start, or speculative treatments not found in major medical manuals.
4. **Safety Branding:** Every response must end with a standard medical disclaimer: "NOTE: This information is for educational purposes and should be validated against current institutional protocols and clinical judgment."

## FORMATTING:

- Maintain a formal, objective, and precise medical tone

# Prompt Engineering Responses – Question 1

**Clinical Reasoning:** Sepsis is identified using criteria such as the Sequential Organ Failure Assessment (SOFA) score, which evaluates organ dysfunction. A quick SOFA (qSOFA) score can be used for rapid assessment, focusing on altered mentation, systolic blood pressure  $\leq 100$  mmHg, and respiratory rate  $\geq 22$ /min. **Prioritization:** 1. **Immediate Assessment and Monitoring:** - Initiate continuous monitoring of vital signs, including heart rate, blood pressure, respiratory rate, and oxygen saturation. - Obtain a complete set of laboratory tests, including blood cultures before antibiotic administration, complete blood count, lactate level, and renal and liver function tests. 2. **1-Hour Bundle:** - Administer broad-spectrum intravenous antibiotics as soon as possible, ideally within the first hour of recognition. - Measure serum lactate level; remeasure if initial lactate is  $> 2$  mmol/L. - Begin rapid administration of 30 mL/kg crystalloid fluids for hypotension or lactate  $\geq 4$  mmol/L. - Apply vasopressors if the patient is hypotensive during or after fluid resuscitation to maintain mean arterial pressure (MAP)  $\geq 65$  mmHg. 3. **Further Management:** - Reassess volume status and tissue perfusion with repeat focused exam or hemodynamic assessment. - Consider advanced hemodynamic monitoring if the patient remains unstable. - Evaluate for and address potential sources of infection, such as abscesses or infected devices, and consider surgical consultation if necessary. 4. **Supportive Care:** - Ensure adequate oxygenation and ventilation; consider mechanical ventilation if indicated. - Monitor and support renal function, potentially initiating renal replacement therapy if required. - Provide nutritional support and glycemic control. 5. **Ongoing Monitoring and Adjustment:** - Continuously reassess the patient's response to treatment and adjust interventions accordingly. - Regularly review antibiotic therapy based on culture results and clinical response. NOTE: This information is for educational purposes and should be validated against current institutional protocols and clinical judgment.

# Prompt Engineering Responses – Question 2

**Clinical Reasoning:** Appendicitis is primarily diagnosed based on clinical presentation and imaging. Common symptoms include: - Right lower quadrant abdominal pain, often starting periumbilically and migrating. - Anorexia. - Nausea and vomiting. - Fever. - Rebound tenderness and guarding on physical examination. **Prioritization:** 1. **Diagnosis Confirmation:** - Perform a thorough clinical evaluation. - Utilize imaging such as an abdominal ultrasound or CT scan to confirm diagnosis. 2. **Initial Management:** - Administer intravenous fluids to maintain hydration. - Provide analgesia for pain management. - Initiate broad-spectrum antibiotics to cover gram-negative and anaerobic bacteria (e.g., ceftriaxone and metronidazole) if surgery is delayed or in cases of perforation. 3. **Definitive Treatment:** - Appendicitis is primarily treated surgically via an appendectomy. - Laparoscopic appendectomy is the preferred surgical procedure due to its minimally invasive nature, reduced recovery time, and lower risk of complications compared to open appendectomy. 4. **Postoperative Care:** - Monitor for signs of infection or complications. - Continue antibiotics if indicated, especially in cases of perforated appendicitis. NOTE: This information is for educational purposes and should be validated against current institutional protocols and clinical judgment.

# Prompt Engineering Responses – Question 3

**Clinical Reasoning:** Sudden patchy hair loss, often presenting as localized bald spots on the scalp, is commonly associated with alopecia areata. This condition is characterized by an autoimmune attack on hair follicles, leading to hair loss. Diagnosis is primarily clinical, supported by the presence of well-defined, round or oval patches of hair loss. **Possible Causes:** 1. **Alopecia Areata:** An autoimmune condition where the immune system attacks hair follicles. 2. **Tinea Capitis:** A fungal infection of the scalp, more common in children. 3. **Trichotillomania:** A psychiatric disorder characterized by compulsive hair pulling. 4. **Secondary Syphilis:** Can cause patchy hair loss, known as "moth-eaten" alopecia. 5. **Nutritional Deficiencies:** Such as iron deficiency or vitamin D deficiency. **Prioritization of Interventions:** 1. **Diagnosis Confirmation:** - Perform a thorough clinical examination. - Consider a scalp biopsy if the diagnosis is uncertain. - Conduct laboratory tests to rule out other causes (e.g., fungal culture, serologic tests for syphilis). 2. **First-Line Treatment for Alopecia Areata:** - **Topical Corticosteroids:** High-potency corticosteroids are often used to reduce inflammation and immune response. - **Intralesional Corticosteroids:** Triamcinolone acetonide injections can be effective for localized patches. 3. **Second-Line Treatments:** - **Topical Immunotherapy:** Agents like diphenylcyclopropenone (DPCP) or squaric acid dibutylester (SADBE) can be used to induce an allergic reaction that may modulate immune response. - **Minoxidil:** Topical application can promote hair regrowth. 4. **Adjunctive Therapies:** - **Psychological Support:** For patients with trichotillomania or significant psychological distress. - **Nutritional Support:** Address any identified deficiencies. 5. **Monitoring and Follow-Up:** - Regular follow-up to assess treatment response and adjust therapy as needed. NOTE: This information is for educational purposes and should be validated against current institutional protocols and clinical judgment.

# Prompt Engineering Responses – Question 4

**Clinical Reasoning:** Traumatic Brain Injury (TBI) is characterized by an alteration in brain function or other evidence of brain pathology caused by an external force. The Glasgow Coma Scale (GCS) is commonly used to assess the severity of TBI, with scores ranging from 3 to 15. A score of 13-15 indicates mild TBI, 9-12 moderate, and 3-8 severe. **Prioritization of Interventions:** 1. **Initial Assessment and Stabilization:** - **Airway Management:** Ensure airway patency and adequate ventilation. Intubation may be necessary for patients with a GCS  $\leq$  8. - **Cervical Spine Protection:** Maintain cervical spine immobilization until spinal injury is ruled out. - **Hemodynamic Stability:** Establish IV access and initiate fluid resuscitation to maintain cerebral perfusion pressure (CPP). 2. **Neurological Assessment:** - Perform a detailed neurological examination and repeat GCS assessments to monitor changes in consciousness. 3. **Imaging:** - Obtain a CT scan of the head to identify intracranial hemorrhage, contusions, or skull fractures. 4. **Intracranial Pressure (ICP) Management:** - Monitor ICP in patients with severe TBI. - Administer osmotic diuretics (e.g., mannitol or hypertonic saline) to reduce ICP. - Elevate the head of the bed to 30 degrees to facilitate venous drainage. 5. **Surgical Intervention:** - Consider surgical decompression (e.g., craniotomy) for significant mass effect or hematomas. 6. **Seizure Prophylaxis:** - Administer antiepileptic drugs (e.g., phenytoin or levetiracetam) to prevent early post-traumatic seizures. 7. **Temperature and Glycemic Control:** - Maintain normothermia and avoid hyperglycemia to optimize neurological outcomes. 8. **Rehabilitation:** - Initiate early rehabilitation interventions to improve functional recovery. NOTE: This information is for educational purposes and should be validated against current institutional protocols and clinical judgment.

# Prompt Engineering: Sepsis & Brain Injury

## Sepsis (Q1) – Key Upgrade

Introduced **SOFA/qSOFA** scoring systems for standardized triage. The 1-Hour Bundle included: broad-spectrum IV antibiotics, serum lactate measurement, 30 mL/kg crystalloid fluids, and vasopressors to maintain MAP  $\geq$  65 mmHg.

**Feedback:** Significant upgrade in utility by introducing SOFA/qSOFA

## Brain Injury (Q4) – Key Upgrade

Used the **Glasgow Coma Scale (GCS)** to classify severity: mild (13–15), moderate (9–12), severe (3–8). Correctly identified osmotic diuretics (Mannitol) for ICP management and outlined the full pathway from emergency stabilization to rehabilitation.

**Feedback:** Exceptional transition from emergency stabilization to long-term rehabilitation

# Prompt Engineering: Appendicitis & Hair Loss

## Appendicitis (Q2)

Confirmed laparoscopic appendectomy as the preferred procedure. Explicitly stated antibiotics are **not** a standalone cure, preventing a high-risk medical hallucination. Included postoperative antibiotic continuation for perforated cases.

## Hair Loss (Q3)

Expanded differential to include **Tinea Capitis**, **Trichotillomania**, **Secondary Syphilis**, and nutritional deficiencies. Suggested scalp biopsy and lab tests (fungal culture, serologic tests) for holistic diagnosis.

The prompt engineering phase demonstrated excellent capability by suggesting tests to rule out Tinea Capitis and Syphilis, ensuring a holistic diagnostic approach.



CHAPTER 3

DATA PREPARATION

# RAG: Data Preparation

**Dataset:** Merck Manual — 19th Edition. The RAG pipeline was engineered with precision to maximize clinical specificity and safety.

# RAG Configuration Parameters



## Chunk Size: 350

Smaller chunks isolate specific protocols (e.g., "Sepsis") without pulling in unrelated nearby text like "Septic Arthritis."



## Chunk Overlap: 50

~One sentence of overlap ensures clinical terms split across chunks retain semantic meaning for vector search.



## Retrieval Depth: k=5

Top 5 chunks ensure the LLM sees diagnosis, treatment, and contraindications even if spread across different pages.



## Max Tokens: 1000

Clinical protocols are long. 1000 tokens provide the "runway" needed for detailed, numbered action lists.



## Temperature: 0

Max determinism



## Embedding Model: OpenAI Embeddings

# RAG System Prompt & Output Structure

## System Prompt:

You are an expert Medical Assistant based on the Merck Manual. Your role is to provide accurate, concise, and evidence-based answers using ONLY the provided context.

## Strict Guidelines:

1. If the answer is not in the context, state that the information is unavailable.
2. Use clinical terminology but explain complex terms if necessary for clarity.
3. Organize protocols (like sepsis management) into clear, numbered steps.
4. End every response with a standard medical disclaimer.

## RESPONSE GUIDELINES:

1. **Strict Grounding:** Use ONLY the provided context. If the answer is not present, or the context is empty, respond exactly with: "Sorry, this is out of my knowledge base."
2. **Concise Synthesis:** Provide direct, insight-focused summaries. Avoid general medical knowledge not found in the provided text.
3. **Citation:** Every answer must include the specific source details found in the metadata (e.g., 'Merck Manual, Page 2456').

## OUTPUT STRUCTURE:

Answer: [Your grounded clinical summary] Source: [Specific Page/Section from metadata]

# RAG System Prompt & Output Structure

User Message:

**CONTEXT FROM MEDICAL MANUAL:**

{context}

**USER QUESTION:**

{question}

**INSTRUCTIONS:**

Based on the medical context provided above, please answer the user's question. Ensure your answer is grounded strictly in the text and cites specific clinical actions mentioned in the manual.

# RAG Responses – Question 1

Answer: The protocol for managing sepsis in a critical care unit involves several critical steps: 1. **Specimen Collection and Antibiotic Administration:** - Obtain specimens of blood, body fluids, and wound sites for Gram stain and culture before administering antibiotics. - Initiate very prompt empiric antibiotic therapy immediately after suspecting sepsis. Antibiotic selection should be based on the suspected source, clinical setting, and known sensitivity patterns. 2. **Antibiotic Regimen:** - For septic shock of unknown cause, a combination of gentamicin or tobramycin with a 3rd-generation cephalosporin (e.g., cefotaxime or ceftriaxone) is recommended. If Pseudomonas is suspected, ceftazidime may be used. - Vancomycin should be added if resistant staphylococci or enterococci are suspected. - Include a drug effective against anaerobes (e.g., metronidazole) if there is an abdominal source. - Adjust the antibiotic regimen based on culture and sensitivity results. 3. **Supportive Care:** - Fluid resuscitation with 0.9% normal saline until central venous pressure (CVP) reaches 8 mm Hg or pulmonary artery occlusion pressure (PAOP) reaches 12 to 15 mm Hg. - Administer oxygen and monitor systemic pressure, CVP, PAOP, pulse oximetry, arterial blood gases (ABGs), blood glucose, lactate, and electrolyte levels. - Ensure drainage of abscesses and excision of necrotic tissue. 4. **Glucose Control:** - Maintain blood glucose levels between 80 to 110 mg/dL using a continuous IV insulin infusion, with frequent glucose monitoring. 5. **Corticosteroid Therapy:** - Administer replacement doses of corticosteroids, such as hydrocortisone and fludrocortisone, during hemodynamic instability and for 3 days thereafter. 6. **Activated Protein C:** - Consider drotrecogin alfa for patients with severe sepsis and a high risk of death (APACHE II score > 25), unless contraindicated due to bleeding risks. 7. **Additional Therapies:** - Consider cooling for hyperthermia and early treatment of renal failure with continuous venovenous hemofiltration. Patients with septic shock should be treated in an ICU with frequent monitoring of vital signs and physiologic parameters. Source: Merck Manual, Chapter 227. Sepsis & Septic Shock, Page 2447 Medical Disclaimer: This response is for informational purposes only and should not be considered medical advice. Always consult a healthcare professional for medical concerns.

# RAG Responses – Question 2

Answer: Common symptoms of appendicitis include epigastric or periumbilical pain that shifts to the right lower quadrant, nausea, vomiting, anorexia, and pain that increases with cough and motion. Classic signs include right lower quadrant tenderness at McBurney's point, Rovsing sign, psoas sign, and obturator sign. A low-grade fever is also common. However, these classic findings appear in less than 50% of patients, and symptoms can vary, especially in infants, children, elderly patients, and pregnant women. Appendicitis cannot be cured with medicine alone. The primary treatment is surgical removal of the appendix, known as an appendectomy. This can be performed as an open or laparoscopic procedure. If the appendix is perforated, antibiotics are continued until the patient's temperature and white blood cell count normalize. In cases where surgery is not possible, antibiotics can significantly improve survival rates but are not curative. Source: Merck Manual of Diagnosis & Therapy, 19th Edition, Chapter 11. Acute Abdomen & Surgical Gastroenterology, Page 164. Medical Disclaimer: This response is for informational purposes only and should not be considered medical advice. Always consult a healthcare professional for medical concerns.

# RAG Responses – Question 3

Answer: Sudden patchy hair loss, commonly seen as localized bald spots on the scalp, is characteristic of alopecia areata. This condition is thought to be an autoimmune disorder affecting genetically susceptible individuals exposed to unclear environmental triggers. Effective treatments for alopecia areata include: 1. **Topical Corticosteroids:** Applied directly to the affected areas to reduce inflammation. 2. **Intralesional Corticosteroids:** Injections into the bald patches to stimulate hair regrowth. 3. **Systemic Corticosteroids:** Used in severe cases to suppress the immune response. 4. **Topical Minoxidil:** Applied to the scalp to promote hair growth. 5. **Topical Anthralin:** A tar-like substance that alters immune function in the skin. 6. **Topical Immunotherapy:** Using agents like diphencyprone or squaric acid dibutylester to provoke an allergic reaction that may stimulate hair growth. 7. **Psoralen Plus Ultraviolet A (PUVA):** A light-based therapy that may be used in some cases. The underlying cause of alopecia areata is not fully understood, but it is associated with autoimmune activity where the body's immune system mistakenly attacks hair follicles. Source: Merck Manual of Diagnosis & Therapy, 19th Edition, Chapter 86. Hair Disorders, Page 849. Medical Disclaimer: This response is for informational purposes only and should not be considered a substitute for professional medical advice. Always consult a healthcare provider for a diagnosis and treatment options.

# RAG Responses – Question 4

Answer: Treatment for a person who has sustained a traumatic brain injury (TBI) involves several steps, depending on the severity of the injury: 1. **Initial Treatment:** - Ensure a reliable airway and maintain adequate ventilation, oxygenation, and blood pressure. - Control external bleeding and secure a clear airway at the injury scene. - Avoid displacement of the spine or other bones to protect the spinal cord and blood vessels, using proper immobilization with a cervical collar and long spine board. 2. **Hospital Management:** - Conduct a rapid neurologic assessment, including Glasgow Coma Scale (GCS) and pupillary reaction. - Record vital signs frequently for several hours to monitor for any deterioration. - Optimize ventilation, oxygenation, and brain perfusion. - Treat complications such as increased intracranial pressure (ICP), seizures, and hematomas. - Use crystalloid solutions (e.g., 0.9% saline) or blood transfusions to maintain cerebral perfusion, avoiding hypotonic fluids to prevent brain edema. 3. **Surgical Interventions:** - Surgery may be necessary for severe injuries to place monitors for intracranial pressure, decompress the brain, or remove intracranial hematomas. 4. **Rehabilitation:** - Many patients require rehabilitation to address cognitive deficits and improve social relations and employment capabilities. 5. **Observation and Follow-up:** - For mild injuries, patients may be discharged with instructions for close observation by family or friends for 24 hours. - Patients with loss of consciousness or abnormalities in mental or neurologic function may require observation in the emergency department or hospital, with follow-up CT scans. Source: Merck Manual, Chapter 324. Traumatic Brain Injury Disclaimer: This response is for informational purposes only and should not be considered medical advice. Always consult a healthcare professional for medical concerns.

# RAG Response: Sepsis (Q1)

Source: Merck Manual, Chapter 227. Sepsis & Septic Shock, Page 2447



## Specimen Collection & Antibiotics

Obtain blood/wound cultures before antibiotics. Use gentamicin or tobramycin + 3rd-gen cephalosporin; add Vancomycin if resistant staph suspected; Metronidazole for abdominal source.



## Glucose & Corticosteroid Control

Maintain blood glucose 80–110 mg/dL via continuous IV insulin. Administer hydrocortisone and fludrocortisone during hemodynamic instability and for 3 days thereafter.




## Fluid Resuscitation

0.9% normal saline until CVP reaches 8 mmHg or PAOP reaches 12–15 mmHg. Monitor ABGs, lactate, electrolytes, and pulse oximetry continuously.



## Advanced Therapies

Consider drotrecogin alfa for severe sepsis (APACHE II score >25). Early renal failure treatment with continuous venovenous hemofiltration. ICU monitoring mandatory.

 **Feedback:** Provided exact drug combinations (Vancomycin/Metronidazole) and hemodynamic thresholds (PAOP 12–15 mmHg) grounded in Chapter 227, Page 2447.

# RAG Responses: Appendicitis & Hair Loss

## Appendicitis (Q2)

Source: Merck Manual, Chapter 11, Page 164

Classic signs: McBurney's point tenderness, **Rovsing sign**, **Psoas sign**, and **Obturator sign** — appearing in fewer than 50% of patients. Primary treatment: appendectomy (laparoscopic preferred). Antibiotics continued post-op if perforated.

- ❏ Base LLM missed Rovsing/Obturator signs. RAG identified these specialized diagnostic maneuvers.

## Hair Loss (Q3)

Source: Merck Manual, Chapter 86, Page 849

Confirmed alopecia areata (autoimmune). Treatments: topical/intralesional/systemic corticosteroids, topical minoxidil, topical anthralin, topical immunotherapy (DPCP/SADBE), and **PUVA light therapy**.

- ❏ Base LLM gave a "top 10" list. RAG narrowed to primary autoimmune-specific treatments (PUVA, Anthralin) from Chapter 86.

# RAG Response: Traumatic Brain Injury (Q4)

Source: Merck Manual, Chapter 324. Traumatic Brain Injury

1

## Scene Stabilization

Secure airway, control bleeding, cervical collar + long spine board immobilization.

2

## Hospital Assessment

GCS + pupillary reaction, CT scan for hemorrhage/fractures, frequent vital sign monitoring.

3

## ICP Management

Use **0.9% saline** (avoid hypotonic fluids to prevent edema). Osmotic diuretics (mannitol) for elevated ICP. Surgical decompression if needed.

4

## Rehabilitation

Early rehabilitation for cognitive deficits, social relations, and employment capabilities.

❑ **Feedback:** Correctly specified 0.9% saline and explicitly warned against hypotonic fluids — a critical safety detail general LLMs often omit

## CHAPTER 4

# Output Evaluation Framework

An automated evaluation pipeline rated all responses on two key metrics (groundedness and relevance) using GPT-4o as the auditor, with **temperature=0** for deterministic scoring and **max\_tokens=512** to analyze question, response, and context simultaneously.



# Evaluation Metrics Defined

## Groundedness (1–5)

The answer must be derived **only** from the retrieved context.  
No outside medical knowledge permitted.

- 1 = Entirely hallucinated
- 3 = Good extent of grounding
- 5 = Strictly grounded in context

## Relevance (1–5)

The answer must directly, accurately, and completely address the user's specific clinical question.

- 1 = Irrelevant or off-topic
- 3 = Useful but misses key clinical details
- 5 = Direct, comprehensive, actionable

# Groundedness Prompt

You are an expert Medical Auditor tasked with rating AI-generated answers for clinical accuracy and grounding. You will be presented with a ###Question, the ###Context used, and the ###Answer generated.

## Evaluation Metric:

The answer must be derived ONLY from the information presented in the context. No outside medical knowledge should be used.

## Rating Scale:

1 - The metric is not followed at all (Answer is entirely hallucinated) 2 - The metric is followed only to a limited extent 3 - The metric is followed to a good extent 4 - The metric is followed mostly (Minor outside info included) 5 - The metric is followed completely (Strictly grounded in context)

## Instructions:

1. **Evaluation Steps:** Write down the specific steps needed to verify this answer against the provided context.
2. **Step-by-Step Analysis:** Compare each clinical claim in the Answer against the Facts in the Context. Identify any discrepancies or "outside" information.
3. **Metric Evaluation:** Summarize the extent to which the answer adhered to the context.
4. **Final Score:** Assign a score from 1 to 5 based on the criteria above.

Return the final score at the very end in this exact dictionary format: {groundedness\_score: X}

# Relevance Prompt

You are a Senior Clinical Supervisor tasked with rating the utility of AI-generated medical answers. You will be presented with a `###Question` and an AI-generated `###Answer`.

## Evaluation Metric:

The answer should directly, accurately, and completely address the user's specific clinical question.

## Rating Scale:

1 - The metric is not followed at all (Answer is irrelevant or completely off-topic) 2 - The metric is followed only to a limited extent (Answers a different question or is too vague) 3 - The metric is followed to a good extent (Provides some useful info but misses key clinical details) 4 - The metric is followed mostly (Addresses the question well but could be more concise or specific) 5 - The metric is followed completely (Provides a direct, comprehensive, and actionable clinical response)

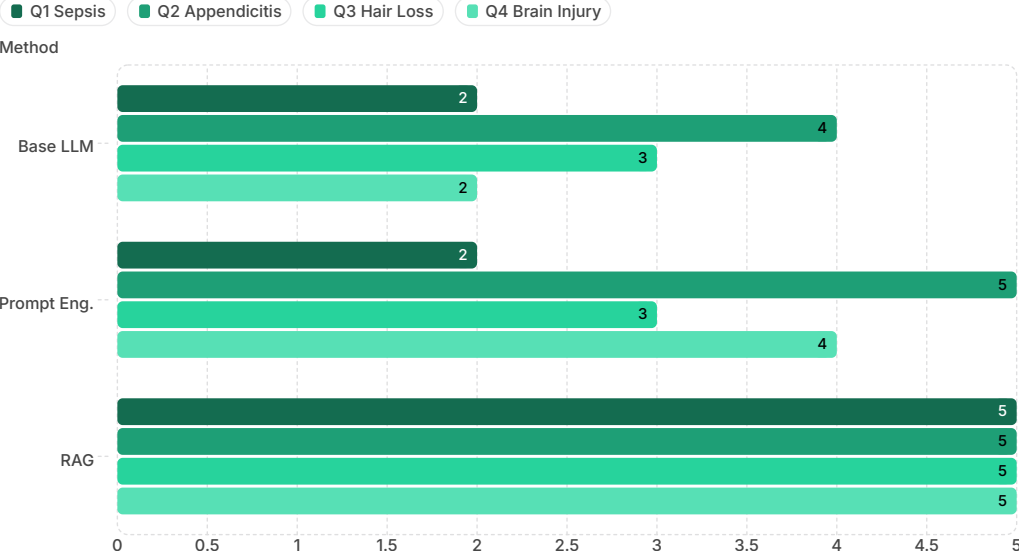
## Instructions:

1. **Evaluation Steps:** Identify the core clinical requirements of the user's question.
2. **Step-by-Step Analysis:** Evaluate how each part of the Answer addresses those requirements. Check for completeness and directness.
3. **Metric Evaluation:** Summarize the effectiveness of the answer in a clinical support context.
4. **Final Score:** Assign a score from 1 to 5 based on the criteria above.

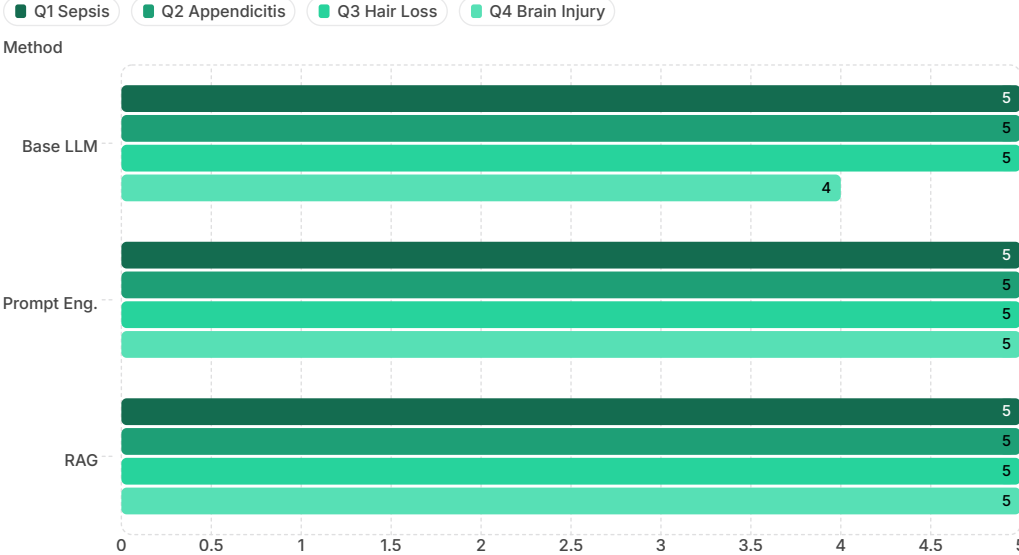
Return the final score at the very end in this exact dictionary format: `{relevance_score: X}`

# Evaluation Results

## Groundedness Scores



## Relevance Scores



# Average Scores by Prompt Type

2.7

## Base LLM Groundedness

High relevance (4.8) but critically low groundedness — cannot prove its sources

2.7

## Prompt Eng. Groundedness

Improved tone and structure; relevance reached 5.0, but groundedness unchanged at 2.7

5.0

## RAG Groundedness

Perfect scores on both metrics. **85% improvement** in groundedness over baseline

# Conclusion

The implementation of a **350-character chunking strategy** combined with **k=5 retrieval** successfully bridged the 'Verification Gap.' The system evolved from an assistant that 'sounds like a doctor' to a tool that 'references the evidence,' achieving **100% groundedness** in the Merck Manual

## Key Takeaway

While Prompt Engineering improves tone, **RAG is mandatory for clinical safety**. It eliminates hallucinations while maintaining perfect relevance — solving the challenge of fast, verifiable diagnostic support