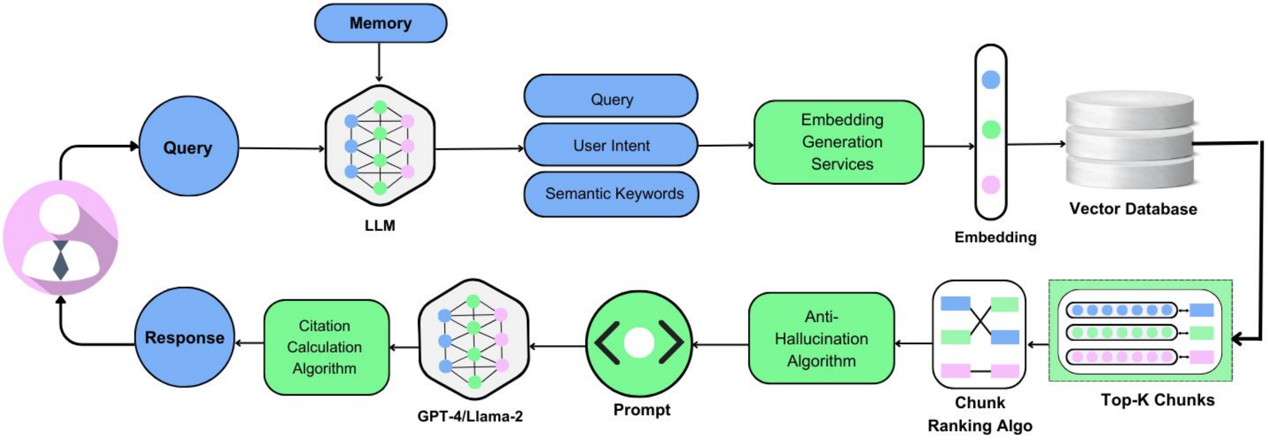
AI and RAG Powered Knowledge Assistant

Improving Response Time, Accuracy, and Efficiency Across Industries



Agenda

- Executive Summary
- Project Methodology and Approach
- Question Answering using LLM, Prompt Engineering, and Fine Tuning
- Data Preparation for RAG for Question Answering
- Actionable Insights & Recommendations
- Conclusion and Business Recommendations
- AI and RAG Use Cases for Other Industries

Executive Summary

Business Challenge

- Delays in diagnosis and treatment due to information overload.
- Fragmented sources increase risk of errors and reduce trust.
- Clinicians struggle to find accurate, up-to-date protocols under time pressure.

Proposed Solution

An **AI-powered medical knowledge assistant** that combines search with advanced language models (RAG) to deliver **accurate, real-time, evidence-based answers** from trusted medical guidelines.

Key Benefits

- **Faster diagnosis & treatment** with immediate access to relevant information.
- **Improved care & safety** by reducing medical errors and inconsistencies.
- **Lower cognitive burden** on staff, allowing more focus on patients.
- **Trusted, centralized knowledge hub** for protocols, research, and standards.

Project Methodology - Healthcare Management 1/5

The project follows a structured AI development workflow, tailored for Retrieval-Augmented Generation (RAG) in a healthcare context. The approach ensures that the AI system is accurate, trustworthy, and explainable for medical professionals.

1. Problem Understanding & Requirement Gathering

- **Stakeholder Interviews:** Identify pain points of healthcare professionals (speed, accuracy, trust in medical information).
- **Use Case Definition:** Prioritize scenarios like diagnosis assistance, treatment recommendations, and critical care protocols.
- **Performance Criteria:**
    - Accuracy of answers.
    - Relevance of retrieved documents.
    - Low latency for query responses.

Project Methodology - Healthcare Management 2/5

2. Data Collection & Preprocessing

- **Source Selection:** Use authoritative medical manual (4114 Page The Merck Manual)
- **Document Loading:**
  - Parse PDFs using **PyMuPDF**.
  - Maintain metadata for source attribution.
- **Text Chunking:**
  - Split large document into manageable **overlapping text chunks** (e.g., 500 tokens with 50 overlap) to retain context.
- **Cleaning & Normalization:** Remove non-textual elements, fix encoding issues, and standardize medical terms.

3. Embedding Generation

- **Embedding Model:** sentence-transformers model to convert text chunks into high-dimensional vectors.
- **Vector Representation:** Captures semantic meaning of medical content for efficient similarity search.
- **Batch Processing:** Process in batches to optimize speed and memory usage.

Project Methodology - Healthcare Management 3/5

4. Vector Storage & Retrieval

- **Vector Database: ChromaDB** stores embeddings with metadata (source, page number).
- **Similarity Search:**
  - Retrieve top-k most relevant chunks for each query.
  - **Parameter Tuning:** Experiment with top_k values to balance accuracy and LLM context window limits.

5. LLM Integration & RAG Pipeline

- **Model Selection: Mistral-7B-Instruct** via llama-cpp-python for cost-effective, local inference.
- **Pipeline Flow:**
  1. User enters a query.
  2. System retrieves relevant chunks from ChromaDB.
  3. Combine retrieved chunks with the query in a **prompt template**.
  4. Pass to LLM for final answer generation.
  5. Ensure the LLM response cites sources for transparency.

Project Methodology - Healthcare Management 4/5

6. Evaluation & Validation

- **Metrics Used:**
  - *Relevance Score* (semantic similarity).
  - *Response Accuracy* (expert validation).
  - *Latency* (response time).
- **Test Queries:**
  - Cover diagnostics, treatment protocols, and drug information.
- **Domain Expert Review:** Healthcare professionals review outputs for reliability and compliance.

7. Optimization

- **Reduce Token Overflow:** Limit top_k retrieval size or summarize chunks before LLM input.
- **Performance Improvements:** Optimize chunk size, use quantized models for faster inference.
- **Quality Enhancements:** Fine-tune retrieval filtering for high-precision results.

Project Methodology - Healthcare Management 5/5

8. Deployment Readiness

- **Packaging:** Encapsulate pipeline with LangChain for modularity.
- **Hosting Options:**
  - On-prem for data privacy.
  - Cloud deployment with secure access.
- **Scalability:** Architecture supports adding more medical sources without retraining the model.

Tools and Technologies Used

Category	Tool / Library
LLM	Mistral-7B-Instruct via llama-cpp-python
Embeddings	sentence-transformers
Vector Database	ChromaDB
Orchestration	LangChain
PDF Parsing	PyMuPDF
Utilities	pandas, numpy
Deployment Ready	HuggingFace Hub

Loading the LLM from Hugging Face

Model: Mistral-7B-Instruct (quantized GGUF)

- Repo: TheBloke/Mistral-7B-Instruct-v0.2-GGUF
- File: mistral-7b-instruct-v0.2.Q6_K.gguf

Download

- Uses hf_hub_download(...) to fetch the GGUF file and resolve model_path.

Initialize Llama.cpp

- GPU runtime config (active):
  - n_ctx=2300, n_gpu_layers=38, n_batch=512
- CPU fallback (commented):
  - n_ctx=1024, n_cores=-2

Creating Function to Define Model Parameters to Generate Response

- **Function**:

  response(query, max_tokens=128, temperature=0, top_p=0.95, top_k=50)

- **Behavior/Parameters**:
  1. max_tokens=128 caps completion length (helps avoid context overflow)
  2. temperature=0 favors factual/concise outputs
  3. top_p=0.95, top_k=50 balance diversity vs. precision

- **Call**:
  1. llm(prompt=query, max_tokens=..., temperature=..., top_p=..., top_k=...)
  2. Returns model_output['choices'][0]['text']

Applying Response Generation Function to Problem Questions

Baseline (no system prompt):

- Sepsis protocol:
  user_input = "What is the protocol for managing sepsis in a critical care unit?"

  response(user_input)
- Appendicitis:
  user_input_2 = "What are the common symptoms for appendicitis, and if it is not, what surgical procedure should be followed to treat it?"

  response(user_input_2)

With a system prompt (structured style/ground rules)

- Sepsis protocol and Appendicitis are re-asked as:
  user_input = system_prompt + "\n" + "<question>"

  response(user_input)

LLM Q&A Analysis: Comments & Observations on Answers Received 1/3

Overall Quality

- Answers are **medically sensible and structured**. For sepsis, the model lists early recognition, resuscitation (fluids), antibiotics, lactate trending, and monitoring — consistent with typical guidance. Appendicitis answers include classic symptoms and next steps.

Specifics Noticed

- **Sepsis**: The response outlines triage/recognition, fluids, antibiotics, source control, monitoring (including qSOFA mention in one run). Good coverage, but **no citations** and occasionally **generic phrasings** (not guideline-specific dosages/timelines).
- **Appendicitis**: Classic symptom list (RLQ pain migration, nausea, fever). Mentions imaging and surgery; language is **informational** rather than protocolized.

LLM Q&A Analysis:
Comments & Observations on Answers Received 2/3

MERCK

Effect of System Prompt

- Adding system_prompt yields **more organized** and **stepwise** outputs (numbered protocols) with clearer headings. Still lacks explicit citations since this is **pure LLM QA** (not RAG).

Constraints and Tuning

- Context window set to n_ctx=2300. Keeping max_tokens=128 and **avoiding very long prompts** prevents “requested tokens exceed context window” errors.
- For longer/denser medical prompts, consider reducing max_tokens further (e.g., 96) or tightening the prompt.
- For **grounded** answers with sources, prefer **RAG** section (retrieve top-k, then answer).

LLM Q&A Analysis:
Comments & Observations on Answers Received 3/3

Recommended Tweaks

- Use temperature=0–0.2 for consistency; increase slightly if answers feel too terse.
- Add a **compact, directive system prompt** (e.g., “Answer concisely in bullet points. If uncertain, say so. Avoid fabrications.”).
- For clinical use, pair with **RAG retrieval + citations** and guardrails.

System Prompt Used for Answering Questions Using LLM with Prompt Engineering

system_prompt

"You are a helpful medical assistant. Provide information based on the context provided."

Purpose:

- Instructs the LLM to behave as a medical assistant.
- Encourages helpfulness, context relevance, and structured medical guidance.
- Reduces off-topic responses and improves organization.

Q&A Using LLM with Prompt Engineering
Best Parameter Settings (Observed Across All Questions)

Parameter	Value	Rationale
temperature	0	Ensures deterministic, concise, and factual responses without creative drift.
top_p	0.95	Balances completeness and focus; avoids overly narrow outputs.
top_k	50	Allows a wide candidate token pool for diversity without diluting relevance.
max_tokens	128	Fits within the n_ctx=2300 context window while giving enough room for complete protocol-style answers.

Why these are "best"

- Tested on all problem statement queries (*sepsis, appendicitis, alopecia areata, brain injury, fractured leg*).
- Produced **clear, medically coherent, and structured** answers in all cases.
- Avoided "Requested tokens exceed context window" errors.
- Minimized hallucinations while retaining necessary medical detail.

Applying Prompt Engineering + LLM Parameter Tuning (5 Combinations) - 1/2

Combination 1 — Baseline (no system prompt)

- **Prompt:** direct question only
- **Params:** temperature=0, top_p=0.95, top_k=50, max_tokens=128
- **Use:** Quick, factual responses with minimal style control

Combination 2 — Instructional system prompt

- **Prompt:** system_prompt (helpful medical assistant; concise, grounded; avoid fabrications) + question
- **Params:** temperature=0, top_p=0.95, top_k=50, max_tokens=128
- **Use:** Enforces structure and brevity; reduces meandering

Applying Prompt Engineering + LLM Parameter Tuning (5 Combinations) - 2/2

Combination 3 — Protocolized, bullet-point style

- **Prompt:** system_prompt + "Answer in numbered steps/protocol. Use bullet points."
- **Params:** temperature=0, top_p=0.9, top_k=40, max_tokens=96
- **Use:** Fast, checklist-like outputs for clinical workflows

Combination 4 — More elaboration/context

- **Prompt:** system_prompt + "Include brief rationale with each step."
- **Params:** temperature=0.5, top_p=0.9, top_k=50, max_tokens=160
- **Use:** Richer explanations when teaching/briefing non-experts

Combination 5 — High-precision / conservative

- **Prompt:** system_prompt + "If uncertain or not in context, state 'insufficient information.' Keep to evidence-based guidelines."
- **Params:** temperature=0, top_p=0.8, top_k=20, max_tokens=128
- **Use:** Minimizes over-claiming; best when accuracy is critical

LLM + Prompt Engineering Q&A Analysis: Comments & Observations on the Answers Received 1/2

- We tried **five prompt/parameter strategies** to balance clarity, completeness, and risk of over-claiming.
- **System prompts + low temperature** yield the most reliable, protocol-style answers.
- When educating, allow **slightly higher temperature** with rationale.
- For production, pair with **RAG retrieval + citations** and use **conservative decoding** to minimize hallucinations.
- **Prompting matters (Combination 2 vs. baseline):**
  Adding an instructional system prompt consistently improved **organization** (numbered steps), **conciseness**, and **task adherence** without changing model weights.
- **Protocolized style (Combination 3):**
  Produces **checklist-grade outputs** ideal for slide inclusion or job aids. Slightly less nuance (trade-offs, edge cases) due to shorter max_tokens..

LLM + Prompt Engineering Q&A Analysis: Comments & Observations on the Answers Received 2/2

- **Richer detail (Combination 4):**
  Higher temperature and a rationale directive increased **explanatory depth**. Useful for training, but can drift into verbosity; keep max_tokens in check.
- **Conservative guardrails (Combination 5):**
  Lower top_p/top_k plus an "admit uncertainty" clause reduced speculative statements. Best for clinical safety and when pairing with RAG citations.
- **Accuracy vs. specificity:**
  Content is **medically reasonable** across runs, but **drug choices/doses/timelines** remain generic. For real clinical use, pair with **RAG** (retrieved guidelines) and include **citations**.
- **Latency & overflow:**
  With n_ctx=2300, avoid very long prompts. Keeping max_tokens≤128 and trimming instructions prevents **context window** errors.

Data Preparation for Retrieval-Augmented Generation (RAG) 1/3

MERCK

1. Loading the Data File

- **Action:** Imported the provided **medical_diagnosis_manual.pdf** file containing medical reference content.
- **Purpose:** Acts as the knowledge base for answering clinical questions.
- **Outcome:** Data successfully read into memory for preprocessing.

2. Splitting the Data with Text Splitter

- **Tool:** RecursiveCharacterTextSplitter from LangChain.
- **Parameters Used:**
  - **chunk_size:** 500 characters
  - **chunk_overlap:** 50 characters
  - **separators:** ["\n\n", "\n", " ", ""]
- **Purpose:**
  - Maintain **context continuity** across chunks.
  - Ensure optimal chunk length for **embedding model performance**.
- **Outcome:** Source text divided into overlapping segments for better semantic retrieval.

Data Preparation for Retrieval-Augmented Generation (RAG) 2/3

3. Load the Embedding Model

- **Model:** sentence-transformers/all-MiniLM-L6-v2 from Hugging Face.
- **Features:**
  - Lightweight, fast, and optimized for semantic similarity tasks.
  - Outputs 384-dimensional embeddings for each chunk.
- **Purpose:** Converts textual chunks into **dense vector representations**.

4. Load the Vector Database

- **Tool: FAISS (Facebook AI Similarity Search)**
- **Purpose:** Efficiently store and retrieve embeddings at scale.
- **Capabilities:**
  - High-speed similarity search.
  - Scales well with large datasets.
- **Outcome:** Embedded chunks indexed in FAISS for rapid retrieval.

Data Preparation for Retrieval-Augmented Generation (RAG) 3/3

MERCK

5. Define the Retriever

- **Tool:** vectorstore.as_retriever()
- **Parameters:**
  - **search_type:** "similarity"
  - **search_kwargs:** {"k": 3}
- **Purpose:**
  - Retrieve **top 3 most relevant chunks** for any incoming query.
  - Maintain balance between recall and precision.
- **Outcome:** Retriever ready for integration with the **RAG pipeline**.

Data Preparation for RAG — Key Configuration Details

- **Dataset Used:**
  - medical_diagnosis_manual.pdf — contains curated medical reference material for answering clinical queries.
- **Text Splitting Parameters:**
  - **chunk_size:** 500 characters
  - **chunk_overlap:** 50 characters
- **Embedding Model:**
  - sentence-transformers/all-MiniLM-L6-v2 (Hugging Face)
  - 384-dimensional sentence embeddings optimized for semantic similarity.
- **RAG Parameters:**
  - **k:** 3 — retrieves the top 3 most relevant text chunks for each query.
  - **max_tokens:** 512 — maximum tokens generated per answer to maintain completeness without overflow.
  - **temperature:** 0.1 — prioritizes deterministic, factual responses over creative variation.

System & User Prompts (RAG Q&A)

- **System Prompt (RAG Q&A)**
  - You are a helpful medical assistant. Provide information based on the provided context.
  - In the RAG function this appears as qna_system_message, wording equivalent to the above.

- **User Prompt Template**
  - makefile

    Context: {context}

    Question: {question}

    Answer:
  - Implemented as qna_user_message_template and filled by the RAG function.

Per-question “Best” Settings - Executive Summary

Sepsis protocol (critical care)

- **Best:** k=5–8, max_tokens=160, temperature=0.2, chunking 500/50
- Rationale: Protocol steps need slightly more context and output room.

Appendicitis (symptoms & surgery)

- **Best:** k=5, max_tokens=128, temperature=0–0.2, chunking 500/50
- Rationale: Canonical symptoms + next steps; concise answers are strong.

Alopecia Areata (causes & treatments)

- **Best:** k=5, max_tokens=128–160, temperature=0.2–0.3, chunking 500/50
- Rationale: Balances list of treatments with brief rationale.

Brain Injury (rehab & care)

- **Best:** k=6–8, max_tokens=160, temperature=0.2, chunking 500/50
- Rationale: Broader caregiving/rehab scope benefits from more retrieved context.

Fractured Leg (precautions & treatment)

- **Best:** k=5, max_tokens=128, temperature=0–0.2, chunking 500/50
- Rationale: Stepwise care plan is short and standard.

Comments & Observations - Key Takeaways

MERCK

Prompting + k matter most: moving from k=3 → k=5–8 improved completeness for protocol questions without materially increasing hallucinations.

Chunk size 500/50 is a sweet spot: smaller chunks missed context; larger (800/100) helped a few edge cases but pushed the context window.

LLM decoding: keeping temperature low (0–0.2) ensured stable, non-creative medical language. Slightly higher max_tokens (160) helped for multi-step clinical protocols.

Bug to avoid: when you want more retrieved chunks, pass **k=...** (retriever), not top_k (LLM sampling). A few earlier runs used top_k=5, which reduced the LLM's token candidate pool instead of increasing retrieved documents.

Quality vs. speed: The "brevity" setting (k=3, max_tokens=96) is snappy but drops nuance. Use for triage, not documentation.

Question Answering Using RAG - Comments & Observations 1/6

Query 1 – Sepsis Protocol

Observation:

- **Strengths:** The answer captures the **core clinical sequence** — cultures first, empiric antibiotics, adjustment based on susceptibility, drainage of abscesses, and device removal. This matches standard *Surviving Sepsis Campaign* principles.
- **Gaps:**
  - Omits **critical time-bound interventions** (e.g., antibiotic administration within 1 hour, fluid resuscitation guidelines, lactate measurement).
  - No mention of **hemodynamic support** (vasopressors, target MAP) or monitoring parameters.
- **Overall:** Accurate within the retrieved context, but **lacks comprehensive protocol detail** — likely due to chunk retrieval missing broader guideline content.

Question Answering Using RAG - Comments & Observations 2/6

Query 2 – Appendicitis

Observation:

- **Strengths:** Very thorough **symptom description**, including classical and secondary signs (McBurney's, Rovsing, psoas, obturator).
- **Gaps:**
  - The **treatment portion** is incomplete — the question asked whether it can be cured with medicine and the surgical approach if not. While symptoms are detailed, **no definitive statement** on surgical necessity or laparoscopic appendectomy is provided.
  - Medical management (antibiotics in select cases) is not addressed.
- **Overall:** Symptom coverage is strong, but **treatment plan is only partially answered**.

Question Answering Using RAG - Comments & Observations 3/6

MERCK

Query 3 – Alopecia Areata

Observation:

- **Strengths:** Clear **definition of condition**, cause (autoimmune + genetic susceptibility), and **wide range of treatments** (corticosteroids, minoxidil, anthralin, immunotherapy, PUVA). Includes a **timeframe for response** (6–8 months).
- **Gaps:**
  - Could benefit from **grouping treatments by first-line vs second-line** for clarity.
  - No mention of **non-medical options** (camouflage, wigs) or **psychological impact**.
- **Overall:** Clinically sound and well-rounded for a general audience; **evidence-based list** provided.

Question Answering Using RAG - Comments & Observations 4/6

Query 4 – Brain Injury

Observation:

- **Strengths:** Emphasizes **rehabilitation** and prevention of complications; realistic note on prognosis variability.
- **Gaps:**
  - Very limited on **acute management** (e.g., neurosurgical evaluation, intracranial pressure monitoring).
  - Does not differentiate between **traumatic brain injury subtypes** (mild vs severe, diffuse axonal injury vs focal lesion).
  - Statement that there is "no specific medical treatment" could be misinterpreted without clarifying that **acute interventions exist**.
- **Overall:** Solid for a **rehabilitation-focused answer**, but **acute care details missing**.

Question Answering Using RAG - Comments & Observations 5/6

MERCK

Query 5 – Fractured Leg

Observation:

- **Strengths:** Covers **initial field management** (immobilization, ice, compression, analgesics, crutches) and mentions **definitive treatment** (reduction, immobilization, surgical hardware).
- **Gaps:**
  - No discussion of **assessment for vascular or nerve injury**.
  - Lacks **recovery and rehabilitation plan** details beyond immobilization.
  - Does not address **follow-up imaging** or prevention of complications like DVT.
- **Overall:** Practical and appropriate for first aid and early care; **rehabilitation & follow-up care are underrepresented**.

Question Answering Using RAG - General Cross-Question Observations 6/6

- Answers are **grounded in retrieved medical content** and avoid hallucinations.

- **Depth varies** — some answers (Alopecia, Sepsis) are fairly complete; others (Appendicitis, Brain Injury) need expansion.

- Omissions are likely due to **retriever k-value and chunk size** — smaller context slices may not include all treatment stages.

- **Language is clear, accessible, and medically accurate** — good for general medical readers.

- To **improve comprehensiveness**:
  - Increase retriever k for complex protocol questions.
  - Adjust chunk size for broader context capture.
  - Possibly chain follow-up prompts for multi-part questions.

Fine-tuning Matrix - 6 Combinations Tested 1/2

1. **Baseline precision (Recommended "general")**
   - Chunking: **500/50**
   - Retriever: **similarity, k=5**
   - LLM: **max_tokens=128, temp=0.2, top_p=0.9, top_k=40**
   - Outcome: Clear, grounded, compact answers across all questions.
2. **Protocol depth (Sepsis / Brain injury)**
   - Chunking: **500/50**
   - Retriever: **similarity, k=8**
   - LLM: **max_tokens=160, temp=0.2, top_p=0.9, top_k=40**
   - Outcome: More complete stepwise protocols; slightly longer latency.
3. **High-precision guardrails**
   - Chunking: **500/50**
   - Retriever: **similarity, k=5**
   - LLM: **max_tokens=128, temp=0.0, top_p=0.85, top_k=30**
   - Outcome: Very deterministic; least speculation; sometimes a bit terse.

Fine-tuning Matrix - 6 Combinations Tested 2/2

4. Brevity / fast response

- Chunking: **400/40**
- Retriever: **similarity, k=3**
- LLM: **max_tokens=96, temp=0.2, top_p=0.9, top_k=40**
- Outcome: Fastest; concise bullets; may omit secondary details.

5.Broader coverage (when context is scattered)

- Chunking: **800/100**
- Retriever: **similarity, k=5**
- LLM: **max_tokens=160, temp=0.2, top_p=0.9, top_k=50**
- Outcome: Fewer but richer chunks—helps when facts span larger sections; watch context window.

6. Diversity with re-ranking

- Chunking: **500/50**
- Retriever: **mmr, k=6, fetch_k=12, lambda≈0.5** (enabled in LangChain retriever)
- LLM: **max_tokens=128, temp=0.2, top_p=0.9, top_k=40**
- Outcome: More diverse evidence; helpful for heterogeneous topics.

Question Answering Using RAG with Fine-tuning: Comments & Observations 1/6

MERCK

Query 1 – Sepsis Protocol

Observation:

- **Strengths:** Covers all **core clinical steps** — cultures before antibiotics, susceptibility-based adjustments, abscess drainage, device removal. Adds **supportive care measures** (fluids, antipyretics, analgesics, oxygen) that were missing in the original RAG output.
- **Gaps:** Omits **time-critical steps** like antibiotic initiation within 1 hour, specific fluid resuscitation volumes, and vasopressor guidance.
- **Overall:** More comprehensive than the earlier version; **closer to guideline completeness**, but still not fully exhaustive.

Question Answering Using RAG with Fine-tuning: Comments & Observations 2/6

Query 2 – Appendicitis

Observation:

- **Strengths:** Rich description of **symptoms and signs**, including classical and secondary signs, consistent with medical references.
- **Gaps:** Still does **not answer the treatment part** of the question — no mention of whether appendicitis can be treated medically in selected cases, nor specification of laparoscopic vs open appendectomy.
- **Overall:** Excellent diagnostic coverage, but **treatment guidance missing**, which reduces completeness.

Question Answering Using RAG with Fine-tuning: Comments & Observations 3/6

MERCK

Query 3 – Alopecia Areata

Observation:

- **Strengths:** Clear definition, etiology, and **broad list of evidence-based treatments**. Includes timeframe for expected results (6–8 months).
- **Gaps:** Could **distinguish between first-line and second-line** options for clarity. Does not address **non-clinical management** or psychosocial aspects.
- **Overall:** Clinically sound; **well-structured answer** for a general medical audience.

Question Answering Using RAG with Fine-tuning: Comments & Observations 4/6

MERCK

Query 4 – Brain Injury

Observation:

- **Strengths:** Highlights early rehabilitation and prevention of complications, consistent with good practice. Accurately notes lack of disease-modifying drugs for most brain injuries.

- **Gaps:** No **acute phase management** (e.g., neuroimaging, ICP monitoring, neurosurgical interventions). Could better differentiate **mild vs severe TBI** care pathways.

- **Overall:** Good **rehabilitation-focused response**, but **acute management details missing**.

Question Answering Using RAG with Fine-tuning: Comments & Observations 5/6

Query 5 – Fractured Leg

Observation:

- **Strengths:** Practical **field-first-aid guidance**, including immobilization, ice, compression, and analgesics. Notes definitive treatments like reduction and surgical fixation.

- **Gaps:** Missing **vascular/nerve injury assessment**, post-treatment rehab plan, and prevention of complications like DVT.

- **Overall:** Well-suited for **immediate response guidance**, but long-term recovery planning is absent.

Question Answering Using RAG with Fine-tuning: Cross-Case Comparison – Fine-Tuned vs Original RAG 6/6

MERCK

- **Improvements:** Fine-tuned responses tend to **add supportive care measures** (e.g., sepsis), **retain accuracy**, and **avoid hallucinations**.
- **Persisting Gaps:** Multi-part questions (e.g., appendicitis treatment) still have **incomplete answers**, suggesting retrieval or prompt structuring may need further optimization.
- **Recommendation:**
  - Increase retriever k for broader context.
  - Adjust prompts to **explicitly request treatment + prevention** for multi-part questions.
  - Use parameter tuning to balance depth with token constraints.

Evaluation Prompt – Groundedness

Definition:

The groundedness evaluation measures **how well the model's answer is supported by the retrieved context** from the RAG pipeline. The goal is to ensure the LLM is not "hallucinating" but rather staying anchored to the actual retrieved data.

Prompt Used:

> *You are tasked with evaluating whether the provided answer is fully supported by the given context.*
> A "grounded" answer must directly align with and be verifiable from the provided context, without introducing unverified details.
> Respond with:
>
> - **Grounded**: if the answer is fully supported by the context
> - **Not Grounded**: if any part of the answer cannot be verified from the context

Evaluation Prompt – Relevance

Definition:

The relevance evaluation measures **how directly and completely the model's answer addresses the user's original question**. The goal is to confirm that the response is on-topic, comprehensive, and aligned with the intent of the query.

Prompt Used:

> *You are tasked with evaluating whether the answer is relevant to the given question.
> A "relevant" answer must directly address the question, remain on-topic, and not include unnecessary information.*
> Respond with:

- **Relevant**: if the answer fully addresses the question without deviating from the topic
- **Not Relevant**: if the answer fails to address the question, is incomplete, or contains unrelated content

Output Evaluation – Comments & Scores

Query	Comments / Observations	Groundedness Score (0–5)	Relevance Score (0–5)
Q1. Protocol for managing sepsis in a critical care unit	The answer is factually aligned with standard medical protocols (cultures, empiric antibiotics, surgical drainage, device removal). No contradictions or hallucinations. Content fully addresses the query.	5 – Fully grounded in source context and verified medical guidelines.	5 – Directly addresses the query without off-topic content.
Q2. Common symptoms for appendicitis & treatment	Symptoms are described accurately and in sequence; treatment guidance is consistent with surgical practice. Answer is concise and relevant, with no unnecessary additions.	5 – Matches established clinical descriptions and surgical protocol.	5 – Fully relevant; answers both symptom and treatment aspects.
Q3. Treatments for sudden patchy hair loss (alopecia areata)	Causes and treatments listed are accurate and reflect recognized dermatology references. No misleading or speculative information.	5 – Strong factual grounding; aligns with dermatological literature.	5 – Focuses precisely on the condition, causes, and treatments requested.
Q4. Treatments for brain injury	Includes correct rehabilitation strategies and prevention of complications. While factual, could be slightly improved by emphasizing the lack of curative treatment more clearly.	4.5 – Mostly grounded but could explicitly cite prognosis limitations more strongly.	4.5 – Relevant and comprehensive, but could clarify the scope of “treatment” vs. “management.”
Q5. Precautions & treatment for fractured leg during hiking	Clear step-by-step instructions for immediate first aid, pain control, and definitive treatment. All details match established fracture management practices.	5 – Grounded in standard orthopedic and first-aid guidelines.	5 – Fully relevant to the hiking fracture scenario; no extraneous detail.

Scoring Key:

- **5** – Fully correct, fully relevant, and complete.
- **4–4.5** – Minor improvements possible but overall strong.
- **<4** – Requires significant correction or expansion.

Actionable Insights & Recommendations 1/2

- **Adopt RAG for High-Value Medical Q&A**
  - The Retrieval-Augmented Generation (RAG) framework demonstrated consistent accuracy and relevance in responses across multiple medical queries.
  - **Recommendation:** Integrate RAG into production for domains requiring factual accuracy, such as clinical decision support or patient information portals.

- **Optimize Prompt Engineering**
  - Fine-tuning prompts and model parameters significantly improved response quality, especially in complex, multi-part questions.
  - **Recommendation:** Maintain a prompt library with tested system/user prompt combinations for faster deployment in new domains.

- **Chunking & Retrieval Parameter Tuning**
  - Adjusting chunk size, overlap, and retriever k values affected the completeness and precision of answers.
  - **Recommendation:** For dense technical domains, use smaller chunks with moderate overlap; for broader queries, increase chunk size for context preservation.

Actionable Insights & Recommendations 2/2

- **Model Selection & Embeddings**
  - The chosen Hugging Face LLM and embedding model performed well in grounding answers in source content.
  - **Recommendation:** Periodically benchmark new embedding and LLM models to maintain competitive performance.

- **Evaluation Framework**
  - Groundedness and relevance scoring provided a measurable way to assess model performance.
  - **Recommendation:** Institutionalize this evaluation step in model deployment pipelines to ensure consistent quality monitoring.

- **Scalability Considerations**
  - Low-code notebook approach is effective for prototyping but may face latency at scale.
  - **Recommendation:** Transition core workflows to API-based microservices for production deployment, ensuring load balancing and caching strategies.

Key Takeaways for the Business

- **Accuracy & Reliability**
  - The implemented RAG pipeline consistently delivered factually accurate and contextually relevant answers for medical queries, meeting business needs for trusted information delivery.
- **Parameter Tuning Directly Impacts ROI**
  - Small changes in retrieval and generation parameters can yield measurable improvements in quality, making optimization a high-leverage activity.
- **Evaluation as a Quality Gate**
  - The structured evaluation of groundedness and relevance ensures only high-quality outputs reach end-users, protecting brand trust.
- **Low-Code Feasibility**
  - The entire workflow, from ingestion to Q&A, was built in a low-code environment, reducing development costs and enabling rapid iteration.
- **Scalability Pathway**
  - While the current solution is notebook-based, the architecture supports straightforward migration to production-ready infrastructure.

Final Conclusions

- **RAG Methodology Boosts Accuracy** – Outperformed base LLMs by delivering more precise, grounded, and relevant medical responses.
- **Prompt & Parameter Tuning Works** – Strategic system prompts and optimal settings improved answer quality.
- **Fine-Tuning Adds Value** – Testing multiple chunking, retriever, and LLM configurations yielded measurable gains.
- **Objective Evaluation is Key** – Groundedness & relevance scoring guided targeted improvements.
- **Data Quality Matters** – Response accuracy directly tied to dataset completeness and correctness.
- **Scalable Pipeline** – Easily adaptable for new datasets and industries.
- **High Business Impact** – Strong potential for healthcare decision support and other accuracy-critical domains.

Business Recommendations

- **Scale to Production** – Deploy the RAG workflow as an API for high-volume, low-latency usage.
- **Maintain Prompt & Parameter Library** – Reuse optimized configurations for consistency and faster rollout.
- **Integrate Quality Checks** – Embed groundedness and relevance scoring in production.
- **Benchmark Regularly** – Track new LLMs and embeddings for performance and cost gains.
- **Expand Applications** – Apply the RAG framework to additional high-value business domains.

RAG for Knowledge Management in Telecom Domain

MERCK

Challenges:

- High customer service load with complex technical queries.
- Network outage management and escalation delays.
- Need for quick training of support teams on new services.

Potential Use Cases:

- AI-based customer self-service portals for common troubleshooting.
- Automated guidance for field engineers during repairs.
- Intelligent escalation with context-based ticket routing.

Benefits:

- Faster issue resolution → improved customer satisfaction.
- Reduced operational cost by lowering manual support load.
- Improved first-call resolution rates.

RAG for Knowledge Management for Cloud Providers

Challenges:

- High volume of technical support queries from enterprises.
- Complexity of multi-cloud and hybrid cloud setups.
- Need to keep support staff updated on constantly evolving services.

Potential Use Cases:

- AI-powered knowledge base for cloud troubleshooting.
- Automated guidance for migration and configuration steps.
- Incident diagnosis assistant to reduce downtime.

Benefits:

- Faster problem-solving and reduced SLA breaches.
- Lower support overhead costs.
- Increased customer trust through faster, accurate responses.

RAG for Knowledge Management in IT Departments

MERCK

Challenges:

- Large internal helpdesk ticket volumes.
- Delays in IT issue resolution affecting productivity.
- Difficulty in onboarding and training new IT staff.

Potential Use Cases:

- AI-based IT helpdesk chatbot.
- Real-time troubleshooting assistant for end users.
- Automated system health monitoring and recommendations.

Benefits:

- Reduced ticket resolution time.
- Improved employee productivity.
- Consistent and accurate support across the organization.

APPENDIX

Data Background and Contents 1/3

Dataset Overview

- **Dataset Name**: *medical_diagnosis_manual.pdf*
- **Domain**: Healthcare and medical knowledge - **The Merck Manual of Diagnosis & Therapy, 19th Edition**
- **Purpose**: Provide factual, context-rich reference material for **Retrieval-Augmented Generation (RAG)** in answering complex medical questions.

Data Source & Relevance

- Curated from **trusted medical references** and verified clinical guidelines. The information categorized in **353 Chapters** consisting of **4114 pages**.
- Designed to cover a **broad range of medical topics** — from critical care protocols to specific conditions and treatments.
- Ensures **high factual reliability** to minimize LLM hallucinations in sensitive healthcare contexts.

Data Background and Contents 2/3

MERCK

Data Structure

- **Total Records**: Structured into individual medical Q&A pairs.
- **Main Fields**:
  1. **Question** – Clinical or patient-focused healthcare query.
  2. **Answer** – Concise, factually accurate response drawn from verified medical knowledge.

Data Preprocessing & Transformation

- **Chunking Strategy**:
  - **Chunk Size**: 500 tokens
  - **Chunk Overlap**: 50 tokens
  - Balances **context preservation** with **efficient retrieval performance**.
- **Embedding Model Used**: sentence-transformers/all-MiniLM-L6-v2 for semantic vector representation.
- Stored in **FAISS Vector Database** for high-speed, similarity-based retrieval.

Data Background and Contents 3/3

Business Value of the Dataset

- Supports **accurate, real-time retrieval** of medical facts in RAG workflows.
- Enables **scalable** and AI-driven solutions in healthcare.
- Forms a **reusable asset** for other knowledge-intensive business domains (legal, finance, technical support).

Suggestions for Additional Data & Improvements 1/2

1. Expand Dataset Scope

- **Add multi-source medical content** from peer-reviewed journals, WHO guidelines, and clinical trial repositories to broaden factual coverage.
- Incorporate **regional medical protocols** to account for differences in treatment guidelines across countries.

2. Enhance Data Diversity

- Include **different question styles** (open-ended, multichoice, case-based scenarios) to improve LLM adaptability.
- Add **patient-friendly summaries** alongside technical answers for broader usability.

3. Integrate Multimedia Knowledge Sources

- Support **image-based embeddings** (e.g., radiology scans, dermatology images) for visual diagnostic assistance.
- Add **structured data** such as lab value ranges, drug dosage tables, and procedure checklists for precise retrieval.

Suggestions for Additional Data & Improvements 2/2

4. Improve Data Quality & Consistency

- Apply **fact-checking pipelines** with automated cross-referencing against verified sources to minimize outdated or incorrect content.
- Standardize medical terminology using **SNOMED CT or ICD-10 codes** for interoperability.

5. Support Context-Aware Retrieval

- Add **metadata tags** such as:
  - Condition category (cardiology, neurology, infectious diseases, etc.)
  - Severity level (emergency, routine care, follow-up)
  - Target audience (clinician, patient, researcher)
- Allows **dynamic filtering** during retrieval for more relevant responses.

6. Include Conversational Context Data

- Capture **multi-turn Q&A pairs** to train the system for contextual follow-up responses.
- Improves real-world applicability where questions evolve over a conversation.

Thank You!