Medical Assistant

NLP with Generative AI

Isaac Gross

Contents / Agenda

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

Executive Summary

Overview:

Our Retrieval-Augmented Generation (RAG) system was evaluated using 5 parameter configurations against 5 medical queries to determine optimal LLM behavior for grounded, relevant clinical responses.

Key Findings:

✅ System prompt and message template significantly improved structure and focus of responses.

✅ Config 5 delivered the most complete, accurate, and clinically aligned answers — scoring 5/5 on both groundedness and relevance for all queries.

⚠️ Earlier configs (1–3) suffered from truncation and less precise grounding due to low max_tokens and default sampling settings.

✅ Ratings from LLM-as-a-Judge (groundedness and relevance) provided objective validation of quality.

Executive Summary

Actionable Insights:

- Use Config 5 parameters for production use in clinical environments:
  - max_tokens = 512
  - temperature = 0.2
  - top_p = 0.9
  - top_k = 30
- Maintain strict system prompt discipline to enforce factual and structured outputs.
- Apply LLM-as-a-Judge framework in future iterations to continuously monitor grounding and quality.

Recommendations:

- Adopt Config 5 as the **baseline for future RAG deployments** in the healthcare domain.
- Continue iterative tuning with edge cases (e.g., rare diseases, ambiguous symptoms).
- Explore fine-tuning or retrieval chunking to enhance granularity and performance on multi-part queries.

Business Problem Overview and Solution Approach

Business Problem:

Healthcare professionals require quick, accurate, and clinically grounded answers to complex medical questions. However, traditional AI assistants often hallucinate, provide incomplete responses, or rely on outdated/non-contextual data. This presents a risk in clinical settings where accuracy is critical.

Challenges:

❌ Inaccurate or hallucinated outputs due to lack of contextual grounding

❌ Incomplete or overly generic responses to nuanced medical questions

❌ Inability to validate whether answers are based on verified sources

Business Problem Overview and Solution Approach

Solution Approach:

To address these challenges, we implemented a Retrieval-Augmented Generation (RAG) pipeline tailored for the medical domain, combined with prompt engineering and systematic parameter tuning to ensure accurate, reliable outputs.

Business Problem Overview and Solution Approach

Methodology:

1. **Contextual Retrieval:** Retrieved relevant context passages using semantic search from a verified medical corpus (Merck Manual).
2. **Prompt Engineering:** Used a strict, domain-specific system prompt and message template to enforce structure, clarity, and factuality.
3. **Parameter Tuning:** Ran 5 test configurations (varying temperature, top_p, top_k, and max_tokens) to optimize performance.
4. **Evaluation Framework:** Used LLM-as-a-Judge to score responses for groundedness and relevance across 5 representative medical queries.

Question Answering using LLM

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

LLM Answer: Starts well with early recognition and symptom list. Cut off after “Resusc…”

Comments / Observations Across 5 Queries:

- ✅ Responses maintained a clinical and factual tone throughout.
- ❌ All responses were truncated mid-way, especially for multi-step medical protocols.
- ✅ Introductory parts of answers were relevant and aligned with medical expectations.
- ❌ Did not fully answer questions involving treatments or procedural names (e.g., appendectomy, rehabilitation).
- ⚠️ Lacked depth and completeness due to low token limit — did not conclude or summarize effectively.
- 💡 Recommendation: Increase max_tokens for longer, more useful answers. Consider adjusting temperature for slightly more flexibility.

Question Answering using LLM

Query: 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?

LLM Answer: Describes symptoms in detail, partially touches on progression of pain.

Comments/Observations:

- ✅ Strong symptom description — clinically aligned.
- ⚠️ Does not answer if it can be treated with medication.
- ❌ Omits surgical treatment name.
- ❌ No conclusion or recommendation — incomplete.
- 💡 Suggest using prompt engineering or RAG to include treatment path explicitly.

Question Answering using LLM

Query: What are the effective treatments or solutions for addressing sudden patchy hair loss, and what could be the possible causes?

LLM Answer: Describes alopecia areata and its autoimmune origins.

Comments/Observations:

- ✅ Accurately identifies condition.
- ✅ Lists possible causes: stress, genetics, infections.
- ❌ No treatment options provided.
- ⚠️ Lacks conclusion or practical advice.
- 💡 RAG needed to extract complete treatment protocols from medical sources.

Question Answering using LLM

Query: What treatments are recommended for a person with traumatic brain injury (TBI)?

LLM Answer: Mentions emergency care and medications, cut off after point 2.

Comments/Observations:

- ✅ Strong start — focuses on emergency care and stabilization.
- ⚠️ Missing key follow-up treatments.
- ❌ Incomplete — cutoff limits usefulness.
- ✅ Clinical tone is appropriate.
- 💡 Increase token count and use retrieval context to include full treatment options.

Question Answering using LLM

Query: What are the necessary precautions and treatment steps for a leg fracture during a hiking trip?

LLM Answer: Mentions calm/stillness, shock symptoms, immobilization — cut off after point 3.

Comments/Observations:

- ✅ Excellent guidance on immediate field precautions.
- ⚠️ Lacks full treatment/recovery steps.
- ❌ Does not mention when/how to evacuate or contact EMS.
- ✅ Tone and structure are helpful for non-medical users.
- 💡 Great candidate for supplementing with RAG-based steps from first aid protocols.

Question Answering using LLM with Prompt Engineering

System Prompt Used:

> *You are a trusted medical assistant. Respond to healthcare-related questions with accurate, clear, and concise medical information. Structure answers in bullet points or numbered steps when applicable. Avoid speculation and only include evidence-based, clinically verified information. Responses should be direct and practical, suitable for use by doctors, nurses, or healthcare students.*

Prompt Design Highlights:

- ✅ **Role definition:** Establishes the assistant as a clinical expert
- ✅ **Clear structure:** Encourages bullet points or procedural steps
- ✅ **Factual accuracy:** Emphasizes verified and evidence-based content only
- ✅ **Audience awareness:** Tailored for use in medical settings or education
- ❌ **No retrieval:** This LLM test is not yet using external documents like the Merck Manual

Question Answering using LLM with Prompt Engineering

Configuration 1 (Baseline)

Parameters Tested:

- max_tokens = 128
- temperature = 0
- top_p = 0.95
- top_k = 50

Comments / Observations Across 5 Queries:

- ✅ Responses improved noticeably with the system prompt — more structured and on-topic.
- ✅ Output followed bullet-point or stepwise format as intended.
- ⚠️ Still experienced truncation due to low max_tokens, especially in complex answers (e.g., sepsis, brain injury).
- ✅ Tone was professional and concise — aligned with medical use cases.
- ❌ Some answers ended abruptly (e.g., sepsis protocol, alopecia treatment, brain rehab), limiting usefulness for end users.
- 💡 Recommendation: Maintain this prompt but increase max_tokens to 300–512 in the next configuration to allow complete answers.
- ⏱️ **Average Response Time:** ~5 seconds per query

Question Answering using LLM with Prompt Engineering

Configuration 2

Parameters Tested:

- max_tokens = 512
- temperature = 0
- top_p = 0.95
- top_k = 50

Comments / Observations Across 5 Queries:

- ✅ Responses were much more complete and detailed compared to the baseline configuration.
- ✅ Output consistently followed structured formats (bullet points, numbered steps), improving readability and clarity.
- ✅ Tone remained professional and medically appropriate, with greater depth in treatment coverage.
- ❌ Some responses were long, which may require truncation or UI constraints for real-world deployment.
- 💡 Prompt paired with extended token limit produced ideal outputs for medical QA use cases.
- ⏱️ **Average Response Time:** ~17.2 seconds per query

Question Answering using LLM with Prompt Engineering

Configuration 3

Parameters Tested:

- max_tokens = 512
- temperature = 0.3
- top_p = 0.9
- top_k = 50

Comments / Observations Across 5 Queries:

- ✅ Responses remained clinically accurate while showing slightly more natural phrasing and explanation.
- ✅ Structure and formatting continued to follow prompt expectations (bullet points, clear headings).
- ✅ Greater detail observed in answers for brain injury and hair loss — including supportive care and alternative therapies.
- ❌ Slightly longer responses in some cases; not all were fully complete (e.g., hair loss cut off at aging).
- 💡 Mild temperature increase introduced helpful flexibility without noticeable hallucination.
- ⏱️ **Average Response Time:** ~15.8 seconds per query

Question Answering using LLM with Prompt Engineering

Configuration 4

Parameters Tested:

- max_tokens = 512
- temperature = 0.7
- top_p = 1.0
- top_k = 40

Comments / Observations Across 5 Queries:

- ✅ Answers maintained structure while expanding with broader context and supportive detail.
- ✅ Tone became slightly more conversational but still aligned with clinical use.
- ✅ Some responses introduced more treatment options (e.g., JAK inhibitors, ECMO), adding helpful nuance.
- ❌ One or two answers ended abruptly again (e.g., hair loss at aging), suggesting length/processing limits.
- 💡 Slight increase in creativity did not compromise accuracy in this test, but continued monitoring for hallucinations is advised.
- ⏱️ **Average Response Time:** ~17.4 seconds per query

Question Answering using LLM with Prompt Engineering

Configuration 5

Parameters Tested:

- max_tokens = 512
- temperature = 0.2
- top_p = 0.9
- top_k = 30

Comments / Observations Across 5 Queries:

- ✅ Responses were complete, factual, and well-structured across all queries.
- ✅ Balanced tone: clear, professional, and concise without being overly clinical or verbose.
- ✅ Strong variety in coverage (e.g., EGDT, corticosteroids, JAK inhibitors, psychological support) while avoiding hallucinations.
- ❌ A few responses leaned heavily into general advice (e.g., Q5 on lifestyle tips), but still relevant to the question.
- 💡 This configuration demonstrated the best trade-off between completeness, precision, and efficiency — ideal for production-level use.
- ⏱️ **Average Response Time:** ~15.6 seconds per query

Question Answering using LLM with Prompt Engineering

Optimal Setup Identified from Testing (Configuration 5):

- max_tokens = 512
- temperature = 0.2
- top_p = 0.9
- top_k = 30

Why This Configuration Was Best:

- ✅ **Complete and accurate responses** across all queries, with no truncation.
- ✅ **Consistently structured outputs** (bullet points, stepwise instructions) aligned with the system prompt.
- ✅ **Professional and clear tone** suitable for clinical or educational use.
- ✅ **Low hallucination risk** — no speculative or unsupported content observed.
- ✅ **Efficient inference** with strong coverage of complex, multi-part questions.

Data Preparation for RAG

📘 Source Document

- **Merck Manual (PDF)**
  Used as a trusted and comprehensive source of medical knowledge.

🔪 Chunking Parameters

- **Chunk Size = 500 tokens**
  → Balances context granularity with model input limits. Large enough to retain medical context but small enough to fit in memory.

- **Chunk Overlap = 75 tokens**
  → Ensures continuity between adjacent chunks. Reduces risk of cutting off key phrases mid-thought, which is important in clinical text.

- **Encoding = 'cl100k_base'**
  → Matches the tokenizer used by most LLMs (including OpenAI and llama.cpp), ensuring accurate token count control.

Data Preparation for RAG

🧠 **Embedding Model**

- **Model: all-mpnet-base-v2**
  → Chosen for **high accuracy** in semantic similarity and general-purpose retrieval tasks, outperforming MiniLM on many benchmarks.

- **Vector Dimension = 768**
  → Defines embedding size. Higher dimensions improve expressiveness for complex contexts like medical text.

🧱 **Vector Store: Chroma**

- **Persist Directory = 'medical_db'**
  → Ensures embeddings are saved and reused between sessions for performance and reproducibility.

Data Preparation for RAG

🔍 Retrieval Settings

- **Search Type = similarity**
  → Retrieves semantically similar chunks using cosine similarity.

- **Top K = 3**
  → Pulls top 3 most relevant chunks per query.
  Chosen to balance:
  ✔ Sufficient context for generation
  ✔ Token limit constraints

Every parameter was chosen to **maximize retrieval quality** while keeping **efficiency and LLM limits** in mind.

Question Answering using RAG

System Prompt Used:

You are a trusted medical assistant. Use the provided context to answer healthcare-related questions with medically accurate, clear, and concise information. Only rely on the supplied context when forming answers — do not use prior knowledge or make assumptions. Structure responses using bullet points or numbered steps when appropriate. Avoid speculation and ensure the content is factual, evidence-based, and clinically verified. Your responses should be direct and practical, intended for use by doctors, nurses, or medical students.

User Prompt Template:

Based on the following context, answer the question:

Context: {context}

Question: {question}

Prompt Design Highlights:

✅ Role definition: Frames the LLM as a clinical assistant.

✅ Context grounding: Forces the model to stay within retrieved chunks.

✅ Structure & clarity: Promotes use of bullet points/steps for better readability.

✅ Factual focus: Avoids hallucinations and reinforces evidence-based reasoning.

❌ No hallucination: Avoids model training data and emphasizes retrieved content only.

Question Answering using RAG

Configuration 1 (Baseline)

Parameters Tested:

- k = 3
- max_tokens = 128
- temperature = 0
- top_p = 0.95
- top_k = 50

Comments / Observations Across 5 Queries:

- ✅ Responses were more structured due to system prompt — tone was professional and medically appropriate.
- ✅ Used bullet points or procedural steps as intended by the prompt template.
- ⚠️ Truncation observed in multi-part or complex answers (e.g., sepsis management, brain injury).
- ❌ Some responses ended abruptly, reducing informativeness for clinical use.
- ⚠️ Retrieval with k=3 was sufficient, but borderline for more nuanced queries.
- ✅ Zero hallucination — model respected the “no fallback” design by not generating unsupported claims.

Question Answering using RAG

Configuration 2

Parameters Tested:

- k = 4
- max_tokens = 512
- temperature = 0.2
- top_p = 0.9
- top_k = 30

Comments / Observations Across 5 Queries:

- ✅ Responses were significantly more complete and detailed — no noticeable truncation.
- ✅ Answers were aligned with the medical context and showed improved recall of multi-step treatments and diagnoses.
- ✅ Tone remained professional and educational, ideal for healthcare learners.
- ✅ Lists and steps were consistently followed, improving clarity and usability.
- ⚠️ Some responses (e.g., Q1, Q4) are still long enough to test token limits — monitor for edge cases.
- ✅ RAG behavior improved due to higher k — more relevant details pulled in across documents.
- ❌ Slight overlap or redundancy in phrasing occasionally observed — consider chunk tuning if persistent.

Question Answering using RAG

Configuration 3

Parameters Tested:

- k = 3
- max_tokens = 384
- temperature = 0.3
- top_p = 0.9
- top_k = 20

Comments / Observations Across 5 Queries:

- ✅ Responses were more concise and clinically specific without losing key detail.
- ✅ Slightly higher `k` value improved supporting context, reducing ambiguity in answers.
- ✅ Increasing `max_tokens` minimized truncation issues seen in previous setup.
- ⚠️ Still saw slight cut-off on long, multi-part questions (e.g., brain injury).
- ✅ Better tone consistency — professional and aligned with the medical assistant persona.
- ⚠️ Some repetitions and formatting drift occurred in longer lists.

Question Answering using RAG

Configuration 4

Parameters Tested:

- k = 4
- max_tokens = 512
- temperature = 0.5
- top_p = 0.9
- top_k = 20

Comments / Observations Across 5 Queries:

- ✅ Answers were more complete and well-rounded, especially in complex multi-step queries like TBI and sepsis protocols.
- ✅ Maintained clinical tone and factual accuracy, aligning with the system prompt.
- ✅ Reduced truncation issues compared to lower token limits in prior configs.
- ⚠️ Some repetition and mild verbosity noted in a few answers (e.g., alopecia and fractures).
- ⚠️ Slightly slower response time due to larger token generation and increased retrieved context.
- ✅ High utility for medical professionals with actionable and structured guidance.

Question Answering using RAG

Configuration 5

Parameters Tested:

- k = 3
- max_tokens = 512
- temperature = 0.3
- top_p = 0.85
- top_k = 20

Comments / Observations Across 5 Queries:

- ✅ Responses were consistently complete and clinically accurate, with no truncation or missing steps.
- ✅ Format followed a clean, stepwise or bulleted structure, ideal for medical reference or educational use.
- ✅ Tone remained professional and direct, with minimized redundancy.
- ✅ Reduced hallucination or repetition compared to Config 4, even on complex topics like brain injury.
- ⚠️ In a few cases, slightly dense answers with long lists (e.g., traumatic brain injury) — could benefit from slight pruning if intended for patient-facing use.
- ✅ Final answers feel well-balanced for coverage, readability, and correctness.

Question Answering using RAG

Optimal Setup Identified from Testing (Configuration 5):

- max_tokens = 512
- temperature = 0.3
- top_p = 0.85
- top_k = 20
- k = 3

Why This Configuration Was Best:

- ✅ **Complete and accurate responses** across all queries — no truncation or loss of detail observed.
- ✅ **Consistently structured outputs** (bullet points, numbered lists) that align well with the system prompt.
- ✅ **Professional and well-aligned tone** for clinical and educational use.
- ✅ **Low hallucination risk** — responses were grounded in the source material with no off-topic or invented content.
- ✅ **Optimized balance** of fluency and precision — handled multi-part queries clearly and thoroughly.

Output Evaluation

Groundedness Evaluation Prompt:

You are a medical domain expert. Your task is to assess whether the assistant’s answer is strictly grounded in the provided context. Do not consider the quality of the answer beyond its alignment with the given context.

Respond with one of the following scores:

5 – Fully grounded

4 – Mostly grounded

3 – Partially grounded

2 – Minimally grounded

1 – Not grounded

Also include a brief justification for your score.

Output Evaluation

Relevance Evaluation Prompt:

You are a medical domain expert. Your task is to rate how well the assistant’s answer addresses the user’s question. Only assess how relevant and complete the response is, not whether it is grounded in context.

Respond with one of the following scores:

5 – Fully relevant

4 – Mostly relevant

3 – Partially relevant

2 – Minimally relevant

1 – Not relevant

Also include a brief justification for your score.

Output Evaluation

Q1 – Sepsis Management

- ⭐ *Groundedness:* 5 – Fully grounded
- ⭐ *Relevance:* 5 – Fully relevant
- ✅ Response thoroughly captured key protocols: empiric antibiotics, glucose control, vasopressors, ICU care. Aligned perfectly with the Merck Manual context.

Q2 – Appendicitis

- ⭐ *Groundedness:* 5 – Fully grounded
- ⭐ *Relevance:* 5 – Fully relevant
- ✅ Accurately identified symptoms and explained surgical intervention as the only treatment, all strictly supported by context.

Output Evaluation

Q3 – Sudden Hair Loss

- ⭐ *Groundedness:* 5 – Fully grounded
- ⭐ *Relevance:* 5 – Fully relevant
- ✅ Clearly attributed alopecia areata as a cause, included medically sound treatments (corticosteroids, immunosuppressants). Contextual alignment was excellent.

Q4 – Traumatic Brain Injury

- ⭐ *Groundedness:* 5 – Fully grounded
- ⭐ *Relevance:* 5 – Fully relevant
- ✅ Strong structure and detail. Covered ventilation, perfusion, intracranial pressure management, rehab, and seizure control — all from the source.

Output Evaluation

Q5 – Leg Fracture Treatment

- ⭐ *Groundedness:* 5 – Fully grounded
- ⭐ *Relevance:* 5 – Fully relevant
- ✅ Comprehensive and clear. Included pain management, immobilization, hygiene, infection risks, and monitoring — all matched to context.